In-Context Learning by Linear Attention: Exact Asymptotics and Experiments

Anonymous Author(s)
Affiliation
Address
email

Abstract

Transformers have a remarkable ability to learn and execute tasks based on examples provided within the input itself, without explicit prior training. It has been argued that this capability, known as in-context learning (ICL), is a cornerstone of Transformers' success, yet questions about the necessary sample complexity, pretraining task diversity, and context length for successful ICL remain unresolved. In this work, we provide precise answers to these questions using a solvable model of ICL for a linear regression task with linear attention. We derive asymptotics for the learning curve in a regime where token dimension, context length, and pretraining diversity scale proportionally, and pretraining examples scale quadratically. Our analysis reveals a double-descent learning curve and a transition between low and high task diversity, which is empirically validated with experiments on realistic Transformer architectures.

1 Introduction

Since their introduction by Vaswani et al. in 2017 [1], Transformers have become a cornerstone of modern artificial intelligence (AI). Transformers achieve state-of-the art performance across many domains, even those that are not inherently sequential [2] as originally intended. Strikingly, they underpin breakthroughs achieved by large language models (LLMs) such as BERT [3], LLaMA [4], and the GPT series [5–8]. The advancements enabled by Transformers have inspired much research aimed at understanding their working principles. One key observation is that LLMs gain new behaviors and skills as their number of parameters and the size of their training datasets grow [7, 9–11]. A particularly important emergent skill is *in-context learning* (ICL), which describes the model's ability to learn and execute tasks based on the context provided within the input itself, without the need for explicit prior training on those specific tasks. ICL enables language models to perform new, specialized tasks without retraining, which is arguably a key reason for their general-purpose abilities.

Despite many recent studies on understanding ICL, important questions about how and when ICL emerges in LLMs are still mostly open. LLMs are trained (or pretrained) with a next token prediction objective. How do the different algorithmic and hyperparameter choices that go into the pretraining procedure affect ICL performance? What algorithms do Transformers implement for ICL? How many pretraining examples are required for ICL to emerge? How many examples should be provided within the input for the model to be able to solve an in-context task? How diverse should the tasks in the training dataset be for in-context learning of truly new tasks not encountered in the training dataset? We address these questions by investigating a simplified model of a Transformer that captures its key architectural motif: the linear self-attention module [12–17]. Linear attention includes the quadratic pairwise interactions between inputs that lie at the heart of softmax attention, but it omits the normalization steps and fully connected layers. This simplification makes the model more amenable to theoretical analysis. Our main result is a sharp asymptotic analysis of ICL for linear regression using linear attention, leading to a more precisely predictive theory than previous population risk analyses or finite-sample bounds [13, 16]. The main contributions of our paper are structured as follows:

We begin in §2 by developing a simplified parameterization of linear self-attention that allows pretraining on the ICL linear regression task to be performed using ridge regression. Within this simplified model, we identify a phenomenologically rich scaling limit in which the ICL performance can be analyzed (§3). In this joint limit, we compute sharp asymptotics for ICL performance using random matrix theory. Our theoretical results reveal several interesting phenomena (§4). First, we observe double-descent in the model's ICL generalization performance as a function of pretraining dataset size, reflecting our assumption that it is pretrained to interpolation. Second, we uncover a transition to in-context learning as the pretraining task diversity increases. This transition recapitulates the empirical findings of [18] in full Transformer models. We further show through numerical experiments that these insights from our theory transfer to full Transformer models with softmax self-attention.

Understanding the mechanistic underpinnings of ICL of well-controlled synthetic tasks in solvable models is an important prerequisite to understanding how it emerges from pretraining on natural data [19].

2 Problem formulation

ICL of linear regression In an ICL task, the model takes as input a sequence of tokens $\{x_1, y_1, x_2, y_2, \dots, x_\ell, y_\ell, x_{\ell+1}\}$, and outputs a prediction of $y_{\ell+1}$. We will often refer to an input sequence as a *context*. We will refer to ℓ as the *context length*. We focus on an approximately linear mapping between $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$:

$$y_i = \langle x_i, w \rangle + \epsilon_i, \tag{1}$$

where ϵ_i is a Gaussian noise with mean zero and variance ρ, and $w \in \mathbb{R}^d$ is referred to as a *task vector*. We note that the task vector w is fixed within a context, but can change between different contexts. The model has to learn w from the ℓ pairs presented within the context, and use it to predict $y_{\ell+1}$ from $x_{\ell+1}$.

Linear self-attention The model that we will analytically study is the linear self-attention block [20]. Linear self-attention takes as input an embedding matrix Z, whose columns hold the sequence tokens. The choice of embedding matrix for a sequence is not unique; here, following the convention in [15, 16, 20], we will embed the input sequence $\{x_1, y_1, x_2, y_2, \dots, x_\ell, y_\ell, x_{\ell+1}\}$ as:

$$Z = \begin{bmatrix} x_1 & x_2 & \dots & x_\ell & x_{\ell+1} \\ y_1 & y_2 & \dots & y_\ell & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(\ell+1)}, \tag{2}$$

where 0 in the lower-right corner is a token that prompts the missing value $y_{\ell+1}$ to be predicted. For appropriately-sized key, query, and value matrices K, Q, V, the output of a linear-attention block [20–22] is given by

$$A := Z + \frac{1}{\ell} V Z (KZ)^\top (QZ). \tag{3}$$

The output A is a matrix while our goal is to predict a scalar, $y_{\ell+1}$. Following the choice of positional encoding in (2), we will take $A_{d+1,\ell+1}$, the element of A corresponding to the 0 prompt, as the prediction for $y_{\ell+1}$, namely $\hat{y} := A_{d+1,\ell+1}$.

Pretraining data The model is pretrained on n sample sequences, where the μth sample is a collection of $\ell + 1$ vector-scalar pairs $\{x_i^\mu \in \mathbb{R}^d, y_i^\mu \in \mathbb{R}\}_{i=1}^{\ell+1}$ related by the approximate linear mapping in (1): $y_i^\mu = \langle x_i^\mu, w^\mu \rangle + \epsilon_i^\mu$. Here, w^μ denotes the task vector associated with the μth sample. We make the following statistical assumptions:

- x_i^μ are d-dimensional random vectors, sampled i.i.d. over both i and μ from $\mathcal{N}(0, I_d/d)$.

- At the start of training, construct a finite set of k elements, written $\Omega_k = \{w_1, w_2, \dots, w_k\}$. The elements of this set are independently drawn once from $w_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$. For $1 \leq \mu \leq n$, the task vector w^μ associated with the μth sample context is uniformly sampled from Ω_k. Note that the variable k controls the task diversity in the pretraining dataset. Importantly, k can be less than n, in which case the same task vector from Ω_k may be repeated multiple times.
- The noise terms ϵ_i^μ are i.i.d. over both i and μ, and drawn from $\mathcal{N}(0, \rho)$.

We denote a sample from this distribution by $(Z, y_{\ell+1}) \sim \mathcal{P}_{\text{train}}$.

Parameter reduction Before specifying a training procedure, we examine the prediction mechanism of the linear attention module for the ICL task. This is a fruitful exercise, shedding light on critical questions: Can linear self-attention learn linear regression in-context? If so, what information do model parameters learn from data in solving this ICL problem?

We start by rewriting the output of the linear attention module $\hat{y} = A_{d+1,\ell+1}$ in an alternative form. Following [16], we define

$$V = \begin{bmatrix} V_{11} & v_{12} \\ v_{21}^\top & v_{22} \end{bmatrix}, \quad M = \begin{bmatrix} M_{11} & m_{12} \\ m_{21}^\top & m_{22} \end{bmatrix} := K^\top Q, \tag{4}$$

where $V_{11} \in \mathbb{R}^{d \times d}$, $v_{12}, v_{21} \in \mathbb{R}^d$, $v_{22} \in \mathbb{R}$, $M_{11} \in \mathbb{R}^{d \times d}$, $m_{12}, m_{21} \in \mathbb{R}^d$, and $m_{22} \in \mathbb{R}$. Expanding (3), one can check that

$$\hat{y} = \frac{1}{\ell} \left\langle x_{\ell+1}, v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i + v_{22} m_{21} \sum_{i=1}^{\ell} y_i^2 + M_{11}^\top \sum_{i=1}^{\ell+1} x_i x_i^\top v_{21} + m_{21} \sum_{i=1}^{\ell} y_i x_i^\top v_{21} \right\rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ stands for the standard inner product.

This expression reveals several interesting points. First, not all parameters in (4) contribute to the output: we can discard all the parameters except for the last row of V and the first d columns of M. Second, the first term

$$\frac{1}{\ell} v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i \tag{6}$$

offers a hint about how the linear attention module might be solving the task. The sum $\frac{1}{\ell} \sum_{i \leq \ell} y_i x_i$ is a noisy estimate of $\mathbb{E}[xx^\top] w$ for that context. Hence, if the parameters of the model are such that $v_{22} M_{11}^\top$ is approximately $\mathbb{E}[xx^\top]^{-1}$, this term alone makes a good prediction for the output. Motivated by this observation, and a more detailed argument presented in Section SI-6 of the Supplementary Information, we study the linear attention module with the constraint $v_{21} = 0$. In this case, we have the model

$$\hat{y} = \langle \Gamma, H_Z \rangle. \tag{7}$$

for

$$\text{Parameter matrix} \quad \Gamma := v_{22} \begin{bmatrix} M_{11}^\top / d & m_{21} \end{bmatrix} \in \mathbb{R}^{d \times (d+1)} \tag{8}$$

$$\text{Input data} \quad H_Z := x_{\ell+1} \begin{bmatrix} \frac{d}{\ell} \sum_{i \leq \ell} y_i x_i^\top & \frac{1}{\ell} \sum_{i \leq \ell} y_i^2 \end{bmatrix} \in \mathbb{R}^{d \times (d+1)}. \tag{9}$$

The $1/d$ scaling of M_{11} in Γ is chosen so that the columns of H_Z scale similarly; it does not affect the final predictor $\hat{y}$.

Model pretraining The parameters of the linear attention module are learned from n samples of input sequences $\{x_1^\mu, y_1^\mu, \dots, x_{\ell+1}^\mu, y_{\ell+1}^\mu\}$ for $\mu = 1, \dots, n$. We estimate model parameters using ridge regression, giving

$$\Gamma^* = \arg\min_{\Gamma} \sum_{\mu=1}^{n} \left(y_{\ell+1}^\mu - \langle \Gamma, H_{Z^\mu} \rangle \right)^2 + \frac{n}{d} \lambda \|\Gamma\|_{\text{F}}^2 , \tag{10}$$

where $\lambda > 0$ is a regularization parameter, and H_{Z^μ} refers to the input matrix (9) populated with the μth sample sequence. The factor n/d in front of λ makes sure that, when we take the $d \to \infty$ or

$n \to \infty$ limits later, there is still a meaningful ridge regularization when $\lambda > 0$. The solution to the optimization problem in (10) can be expressed explicitly as

$$\mathrm{vec}(\Gamma^*) = \left(\frac{n}{d}\lambda I + \sum_{\mu=1}^{n} \mathrm{vec}(H_{Z^\mu})\mathrm{vec}(H_{Z^\mu})^\top \right)^{-1} \sum_{\mu=1}^{n} y_{\ell+1}^{\mu} \mathrm{vec}(H_{Z^\mu}), \tag{11}$$

where $\mathrm{vec}(\cdot)$ denotes the row-major vectorization operation.

Evaluation For a given set of parameters Γ, the model's generalization error is defined as

$$e(\Gamma) \coloneqq \mathbb{E}_{\mathcal{P}_{\mathrm{test}}}\left[\left(y_{\ell+1} - \langle \Gamma, H_Z \rangle \right)^2 \right], \tag{12}$$

where $(Z, y_{\ell+1}) \sim \mathcal{P}_{\mathrm{test}}$ is a new sample drawn from the probability distribution of the test dataset. At test time, x_i and ϵ_i are i.i.d. Gaussians as in the pretraining case. However, for each $1 \le \mu \le n$, the task vector w^μ associated with the μth input sequence is drawn independently from $\mathcal{N}(0, I_d)$. We will denote the test error under this setting by $e^{\mathrm{ICL}}(\Gamma)$.

This ICL task evaluates the true in-context learning performance of the linear attention module. The task vectors in the test set differ from those seen in training, requiring the model to infer them from context. High performance on the ICL task indicates that the model can learn task vectors from the provided context.

To understand the performance of our model on this task, we will need to evaluate these expressions for the pretrained attention matrix Γ^* given in (11). An asymptotically precise prediction of $e^{\mathrm{ICL}}(\Gamma^*)$ will be a main result of this work. We then verify through simulations that the primary insights gained from our theoretical analysis extend to more realistic nonlinear Transformers.

3 Theoretical results

Joint asymptotic limit We have now defined both the structure of the training data as well as the parameters to be optimized. For our theoretical analysis, we consider a joint asymptotic limit in which the input dimension d, the pretraining dataset size n, the context length ℓ, and the number of task vectors in the training set k, go to infinity together such that

$$\frac{\ell}{d} \coloneqq \alpha = \Theta(1), \quad \frac{k}{d} \coloneqq \kappa = \Theta(1), \quad \frac{n}{d^2} \coloneqq \tau = \Theta(1). \tag{13}$$

Identification of these scalings constitutes one of the main results of our paper. As we will see, the linear attention module exhibits rich learning phenomena in this limit.

The intuition for these scaling parameters can be seen as follows. Standard results in linear regression [23–25] show that to estimate a d-dimensional task vector w from the ℓ samples within a context, one needs at least $\ell = \Theta(d)$. The number of unique task vectors that must be seen to estimate the covariance matrix of the true d-dimensional task distribution $\mathcal{N}(0, I_d)$ should also scale with d, *i.e.* $k = \Theta(d)$. Finally, we see from (8) that the number of linear attention parameters to be learned is $\Theta(d^2)$. This suggests that the number of individual contexts the model sees during pretraining should scale similarly, *i.e.*, $n = \Theta(d^2)$.

Learning curves for ICL of linear regression by a linear attention module Our theoretical analysis, explained in detail in the Supplementary Information, leads to an asymptotically precise expression for the generalization error under the ICL test distribution being studied. The exact expressions of this function functions can be found in Section SI-13.2 of the SI. For simplicity, we only present in what follows the ridgeless limit (*i.e.*, $\lambda \to 0^+$) of the asymptotic generalization errors.

Result 1 (ICL generalization error in the ridgeless limit)**.** *Let*

$$q^* \coloneqq \frac{1+\rho}{\alpha}, \quad m^* \coloneqq \mathcal{M}_\kappa(q^*), \quad \mu^* \coloneqq q^* \mathcal{M}_{\kappa/\tau}(q^*), \tag{14}$$

where $\mathcal{M}_\kappa(\cdot)$ is defined in (181) and $\mathcal{M}'_\kappa(\cdot)$ is the derivative of $\mathcal{M}_\kappa(q)$ with respect to q. Then

$$e^{\mathrm{ICL}}_{\mathrm{ridgeless}} \coloneqq \lim_{\lambda \to 0^+} e^{\mathrm{ICL}}(\tau, \alpha, \kappa, \rho, \lambda) \tag{15}$$

$$= \begin{cases} \frac{\tau(1+q^*)}{1-\tau}\left[1 - \tau(1-\mu^*)^2 + \mu^*(\rho/q^* - 1)\right] - 2\tau(1-\mu^*) + (1+\rho) & \tau < 1 \\ (q^*+1)\left(1 - 2q^*m^* - (q^*)^2\mathcal{M}'_\kappa(q^*) + \frac{(\rho + q^* - (q^*)^2 m^*)m^*}{\tau - 1}\right) - 2(1 - q^*m^*) + (1+\rho) & \tau > 1 \end{cases}$$

We derive this result using techniques from random matrix theory. The full setup and technical details are presented in the Supplementary Information in Section SI-9 through Section SI-13. The computations involve analysis of the properties of the finite-sample optimal parameter matrix Γ^*.

4 Observed Phenomena

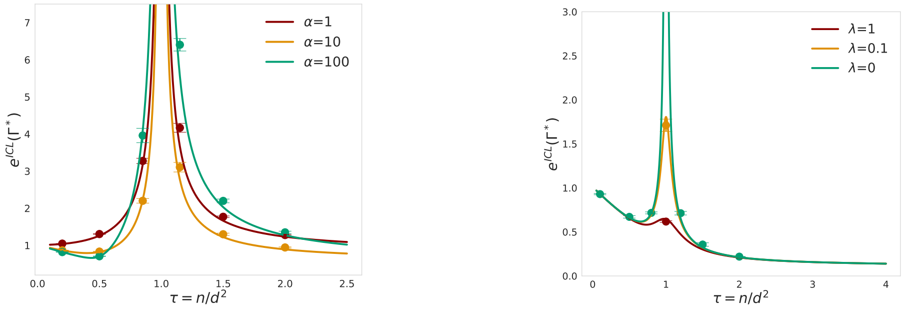
This section discusses two key results that are mathematically evident from our theoretical characterisation of ICL error, namely a double descent in τ and a learning transition in κ. We show how these phenomena follow directly from the theory, and further, remain present in realistic (nonlinear) transformer architectures. A detailed exposition of nonlinear architecture setup and training procedures is given in Section SI-7 in the Supplementary Info. Specific parameter configurations and more detailed descriptions of the figures are available in Section SI-8 in the Supplementary Info.

Double-descent in pretraining samples How large should n, the pretraining dataset size, be for the linear attention to succesfully learn the task in-context? In Figure 1, we plot our theoretical predictions for ICL error as a function of $\tau = n/d^2$ and verify them with numerical simulations. Our results demonstrate that the quadratic scaling of sample size with input dimensions is indeed an appropriate regime where nontrivial learning phenomena can be observed.

As apparent in Figure 1, we find that the generalization error for the ICL task is not monotonic in the number of samples. In the ridgeless limit, ICL error diverges at $\tau = 1$, with the leading order behavior proportional to $(\tau - 1)^{-1}$. This leads to a "double-descent" behavior [25, 26] in the number of samples. As in other models exhibiting double-descent [25–27], the location of the divergence is at the interpolation threshold: the number of parameters of the model (elements of Γ) is, to leading order in d, equal to d^2, which matches the number of pretraining samples at $\tau = 1$. Further, we can investigate the effect of ridge regularisation on the steepness of the double descent, as illustrated in Figure 1b for the ICL task. As we would expect from other models exhibiting double-descent [25–27], increasing the regularization strength suppresses the peak in error around the interpolation threshold.

Figure 2 confirms this phenomenon in a selection of nonlinear models. We recover a peak in error at the interpolation threshold (given by n), and tracking the location of the interpolation threshold as d increases recovers the quadratic scaling $n \sim d^2$.

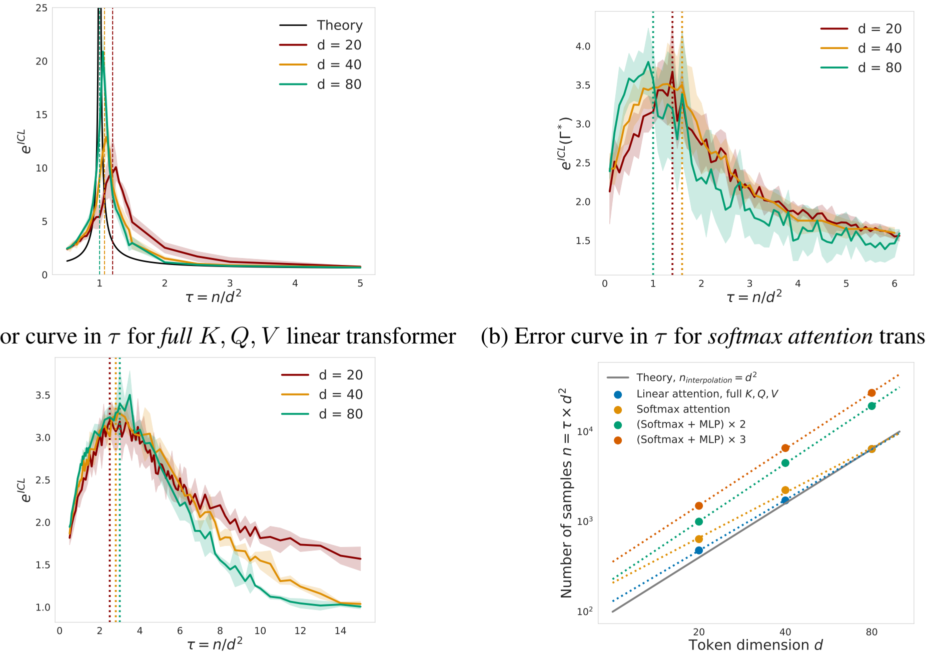
Learning transition with increasing pretraining task diversity Recall that the parameter $\kappa = k/d$ controls the diversity of the training task vectors. How large should it be for ICL to emerge? Figure 3 shows a transition in the performance of a transformer on the ICL task. We see that as κ increases beyond $\kappa = 1$, the ICL error converges rapidly. We interpret this as, in the $\kappa > 1$ regime, the model generalizes to task vectors beyond its pretraining dataset, behaving as if it has learned the true distribution on the task vectors despite having only seen a finite subset in the pretraining dataset. The dependence on α arises since, as α increases, the model achieves even better estimates of the task vector for a single context, allowing it to achieve a better estimate of the true task distribution after



(a) Ridgeless ICL error against τ

(b) Finite ridge ICL error against τ

Figure 1: ICL performance as a function of τ: theory (solid lines) vs simulations (dots). Plots show $e^{\text{ICL}}_{\text{ridgeless}}(\tau, \alpha, \kappa, \rho)$ in 1a and $e^{\text{ICL}}(\tau, \alpha, \kappa, \rho, \lambda)$ in 1b against τ.



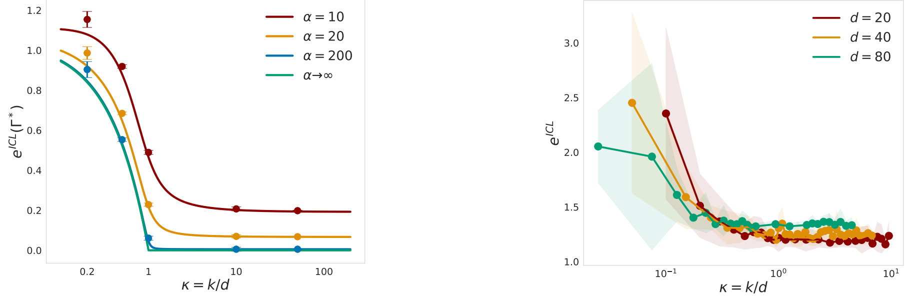
(a) Error curve in τ for *full* K, Q, V linear transformer

(b) Error curve in τ for *softmax attention* transformer

(c) Error curve in τ for transformer w/ two blocks of *softmax attention and 1-layer MLP*

(d) Interpolation threshold follows predicted quadratic $n \propto d^2$ scaling for a range of architectures.

Figure 2: Experimental verification in full linear attention 2a nonlinear models 2b, 2c of both scaling definition for τ and double descent behaviour in n. Figures 2a, 2b, 2c show error curves against τ for various architectures, consistent across token dimension $d = 20, 40, 80$. Double-descent phenomena is confirmed: increasing n will increase error until an interpolation threshold is reached. Coloured dashed lines indicate experimental interpolation threshold for that architecture and d configuration. Figure 2d shows that the location of the interpolation threshold occurs for n proportional to d^2, as predicted by the linear theory. Dots are experimental interpolation thresholds for various architectures, and dashed lines are best fit curves correspond to fitting $\log(n) = a\log(d) + b$, each with $a \approx 2$.



(a) $e^{\mathrm{ICL}}_{\mathrm{ridgeless}}$ against κ for various α and $\tau := 0.2\alpha$

(b) ICL error in non-linear transformer against κ

Figure 3: Plot of transformer generalization error against κ, illustrating sharp transition in performance as pretraining diversity increases. Figure 3a has theory (solid lines) vs simulations (dots). Figure 3b shows ICL performance against κ for the nonlinear architecture given in Figure 2c. This demonstrates consistency of κ scaling across increasing dimension choices $d = 20, 40, 80$, and a similar sharp transition in learning familiar from the linear theory.

seeing multiple contexts with enough task diversity ($\kappa > 1$). Crucially, this learning transition is persistent in nonlinear architectures; an example is seen in Figure 3b.

To explicitly understand the role of κ in the solution learned by the linear attention mechanism, consider the regime where $\tau, \alpha \to \infty$ with $\tau/\alpha = c^*$ kept fixed. Under this setting, we have

$$\lim_{\substack{\tau\to\infty \\ \alpha\to\infty}} e^{\mathrm{ICL}}_{\mathrm{ridgeless}} = \begin{cases} \rho + (1-\kappa)\left(1 + \frac{\rho}{1+\rho}c^*\right) & \kappa < 1 \\ \rho & \kappa > 1 \end{cases}. \tag{16}$$

This change in analytical behavior indicates a phase transition at $\kappa = 1$. Further, the $\kappa > 1$ branch approaches ρ, the information-theoretical error limit for this problem. The smooth learning transitions observed in Figure 3 stem from this phase transition; this behaviour is increasingly obvious for larger α, as can be seen by contrasting the various α curves in Figure 3a.

5 Conclusions

In this work, we compute sharp asymptotics for the in-context learning (ICL) performance in a simplified model of ICL for linear regression using linear attention. This exactly solvable model demonstrates a transition in the generalizing capability of the model as the diversity of pretraining tasks increases, echoing empirical findings in full Transformers [18]. Additionally, we observe a sample-wise double descent as the amount of pretraining data increases. Our numerical experiments show that full, nonlinear Transformers exhibit similar behavior in the scaling regime relevant to our solvable model. Our work represents a first step towards a detailed theoretical understanding of the conditions required for ICL to emerge [19].

In our analysis, we have assumed that the model is trained to interpolation on a fixed dataset. This allows us to cast our simplified form of linear attention pretraining as a ridge regression problem, which in turn enables our random matrix analysis. In contrast, Transformer-based large language models are usually trained in a nearly-online setting, where each gradient update is estimated using fresh examples with no repeating data [28]. Some of our findings, such as double-descent in the learning curve as a function of the number of pretraining examples, are unlikely to generalize to the fully-online setting. It will be interesting to probe these potential differences in future work.

Finally, our results have some bearing on the broad question of what architectural features are required for ICL [7, 11, 19]. Our work shows that a full Transformer—or indeed even full linear attention—is not required for ICL of linear regression. However, our simplified model retains the structured quadratic pairwise interaction between inputs that is at the heart of the attention mechanism. It is this quadratic interaction that allows the model to solve the ICL regression task, which it does essentially by reversing the data correlation. One would therefore hypothesize that our model is minimal in the sense that further simplifications within this model class would impair its ability to solve this ICL task. In the specific context of regression with isotropic data, a simple point of comparison would be to fix $\Gamma = I_d$, which gives a pretraining-free model that should perform well when the context length is very long. However, this further-reduced model would perform poorly if the covariates of the in-context task are anisotropic. More generally, it would be interesting to investigate when models lacking this precisely-engineered quadratic interaction can learn linear regression in-context, and if they are less sample-efficient than the attention-based models considered here.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[9] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

[10] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[11] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD.

[12] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

[13] Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization, 2024.

[14] Pritam Chandra, Tanmay Kumar Sinha, Kabir Ahuja, Ankit Garg, and Navin Goyal. Towards analyzing self-attention via linear neural network, 2024. URL https://openreview.net/forum?id=4fVuBf5HE9.

[15] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

[16] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

[17] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning, 2023.

[18] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

[19] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

[20] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[21] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[22] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[23] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[24] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

[25] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

[26] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

[27] Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[28] Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[30] Sofiia Dubova, Yue M. Lu, Benjamin McKenna, and Horng-Tzer Yau. Universality for the global spectrum of random inner-product kernel matrices in the polynomial regime. *arXiv*, 2023.

[31] Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, Apr 1993. doi: 10.1103/RevModPhys.65.499. URL https://link.aps.org/doi/10.1103/RevModPhys.65.499.

[32] Andreas Engel and Christian van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: https://doi.org/10.1017/CBO9781139164542.

[33] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[34] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4): 667–766, 2022.

[35] Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

[36] Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv:2403.08160*, 2024.

[37] Oussama Dhifallah and Yue M Lu. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

[38] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. *arXiv*, 2024.

[39] Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/montanari22a.html.

[40] László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. The local semicircle law for a general class of random matrices. *Electronic Journal of Probability*, 18(none):1 – 58, 2013. doi: 10.1214/EJP.v18-2473. URL https://doi.org/10.1214/EJP.v18-2473.

[41] László Erdős and Horng-Tzer Yau. *A dynamical approach to random matrix theory*, volume 28. American Mathematical Soc., 2017.

[42] Alston S. Householder. Unitary triangularization of a nonsymmetric matrix. *J. ACM*, 5(4):339–342, oct 1958. ISSN 0004-5411. doi: 10.1145/320941.320947. URL https://doi.org/10.1145/320941.320947.

[43] Yue M. Lu. Householder dice: A matrix-free algorithm for simulating dynamics on Gaussian and random orthogonal ensembles. *IEEE Transactions on Information Theory*, 67(12):8264–8272, 2021. doi: 10.1109/TIT.2021.3114351.

[44] Lloyd N. Trefethen and David Bau, III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997. doi: 10.1137/1.9780898719574. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719574.

Supplementary Information

SI-6 Parameter Reduction

Recall that we can express the output of the linear attention mechanism (with full K, Q, V parameters) as

$$\hat{y} = \frac{1}{\ell}\langle x_{\ell+1}, v_{22}M_{11}^\top \sum_{i=1}^{\ell} y_i x_i + v_{22}m_{21}\sum_{i=1}^{\ell} y_i^2 + M_{11}^\top \sum_{i=1}^{\ell+1} x_i x_i^\top v_{21} + m_{21}\sum_{i=1}^{\ell} y_i x_i^\top v_{21}\rangle, \quad (17)$$

where $\langle \cdot, \cdot \rangle$ stands for the standard inner product. We previously argued that the term

$$\frac{1}{\ell} v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i \quad (18)$$

makes a good prediction for the output. Further, the third term does not depend on outputs y, and thus does not directly contribute to the ICL task that relies on the relationship between x and y. Finally, the last term only considers a one dimensional projection of x onto v_{21}. Because the task vectors w and x are isotropic in the statistical models that we consider, there are no special directions in the problem. Consequently, we expect the optimal v_{21} to be approximately zero by symmetry considerations.

We note that Zhang et al. [16] provide an analysis of population risk (whereas we focus on empirical risk) for a related reduced model in which they set $v_{21} = 0$ and $m_{21} = 0$. Consequently, the predictors they study differ from ours (7) by an additive term. They justify this choice through an optimization argument: if these parameters are initialized to zero, they remain zero under gradient descent optimization of the population risk, given certain conditions.

SI-7 Experimental Details

Our experiments[1] are done with a standard Transformer architecture, where each sample context initially takes the form given by (2). The fully-parameterised linear transformer (fig. 2a) and softmax-only transformer (fig. 2b) do not use MLPs. If MLPs are used (e.g. fig. 2c and fig. 2d), the architecture consists of blocks with: (1) a single-head softmax self-attention with $K, Q, V \in \mathbb{R}^{d+1 \times d+1}$ matrices, followed by (2) a two-layer dense MLP with GELU activation and hidden layer of size $d+1$ [1]. Residual connections are used between the input tokens (padded from dimension d to $d+1$), the pre-MLP output, and the MLP output. We use a variable number of attention+MLP blocks before returning the final logit corresponding to the $(d+1, \ell+1)$th element in the original embedding structure given by (2). The loss function is the mean squared error (MSE) between the predicted label (the output of the model for a given sample Z) and the true value $y_{\ell+1}$. We train the model in an offline setting with n total samples $Z_1, \cdots, Z_n$, divided into 10 batches, using the Adam optimizer [29] with a learning rate 10^{-4} until the training error converges, typically requiring 10000 epochs[2]. The structure of the pretraining and test distributions exactly follows the setup for the ICL task described in Section 2.

SI-8 Figure Details

Figure 1 Simulated errors are calculated by evaluating the corresponding test error on the corresponding optimised Γ^*. *Parameters:* $d = 100$, $\rho = 0.01$ for all; for 1a $\kappa = 0.5$ and for 1b $\alpha = 10, \kappa = \infty$. Averages and standard deviations are computed over 10 runs.

[1] Code to reproduce all experiments will be made available upon acceptance.

[2] Note that larger d models are often trained for less epochs than smaller d models due to early stopping; that said, whether or not early stopping is used in training does not affect either the alignment of error curves in d-scaling nor the qualitative behaviour (double descent in τ and transition in κ).

Figure 2 Interpolation thresholds shown in fig. 2d were computed empirically by searching for location in τ of sharp increase in value and variance of training error at a fixed number of gradient steps. The log-log plot demostrating quadratic scaling of n in d was best-fit on the data points plotted. Explicitly, the exponents of d are $a_{\text{full linear}} = 1.87$, $a_{\text{softmax}} = 1.66$, $a_{2\text{ blocks}} = 2.13$, $a_{3\text{ blocks}} = 2.08$. Theory predicts $a = 2$.

Parameters: $\alpha = 1, \kappa = \infty, \rho = 0.01$. For fig. 2a, 2b, and 2c: variance shown comes from model trained over different samples of pretraining data; lines show averages over 10 runs and shaded region shows standard deviation.

Figure 3 *Parameters for fig. 3a:* $d = 100$, $\tau = 100$. Simulations deviate from theory curve at low κ due to finite size effects. Averages and standard deviations for linear model are computed over 100 runs.

Parameters for fig. 3b: $\tau = 10, \alpha = 1, \rho = 0.01$. Variance shown comes from 10 models trained over different samples of pretraining data.

A note on the subsequent computations

A key technical component of our analysis involves characterizing the spectral properties of the sample covariance matrix of $n = \Theta(d^2)$ i.i.d. random vectors in dimension $\Theta(d^2)$. Each of these vectors is constructed as the vectorized version of the matrix in (9). Related but simpler versions of this type of random matrices involving the tensor product of i.i.d. random vectors have been studied in recent work [30]. Some of our derivations are based on non-rigorous yet technically plausible heuristics. We support these predictions with numerical simulations in the main document and discuss below the steps required to achieve a fully rigorous proof.

Finally, it's worthwhile to comment that this paper and computations therein fall inside a broader program of research that seeks sharp asymptotic characterizations of the performance of machine learning algorithms. This program has a long history in statistical physics [27, 31, 32], and has in recent years attracted substantial attention in machine learning [25, 27, 33–38]. For simplicity, we have assumed that the covariates in the in-context regression problem are drawn from an isotropic Gaussian. However, our technical approach could be extended to anisotropic covariates, and, perhaps more interestingly, to featurized linear attention models in which the inputs are passed through some feature map before linear attention is applied [21, 22]. This extension would be possible thanks to an appropriate form of *Gaussian universality*: for certain classes of regression problems, the asymptotic error coincides with that of a model where the true features are replaced with Gaussian features of matched mean and covariance [25, 30, 33–37, 39]. This would allow for a theoretical characterization of ICL for realistic data structure in a closer approximation of full softmax attention, yielding more precise predictions of how performance scales in real Transformers.

SI-9 Notation

Sets, vectors and matrices: For each $n \in \mathbb{N}$, $[n] \coloneqq \{1, 2, \dots, n\}$. The sphere in $\mathbb{R}^d$ with radius $\sqrt{d}$ is expressed as $\mathcal{S}^{d-1}(\sqrt{d})$. For a vector $v \in \mathbb{R}^d$, its ℓ_2 norm is denoted by $\|v\|$. For a matrix $A \in \mathbb{R}^{d \times d}$, $\|A\|_{\mathsf{op}}$ and $\|A\|_{\mathsf{F}}$ denote the operator (spectral) norm and the Frobenius norm of A, respectively. Additionally, $\|A\|_\infty \coloneqq \max_{i,j \in [n]} \left|A(i,j)\right|$ denotes the entry-wise ℓ_∞ norm. We use e_1 to denote the first natural basis vector $(1, 0, \dots, 0)$, and I is an identity matrix. Their dimensions can be inferred from the context. The trace of A is written as $\operatorname{tr}(A)$.

Our derivations will frequently use the vectorization operation, denoted by $\operatorname{vec}(\cdot)$. It maps a $d_1 \times d_2$ matrix $A \in \mathbb{R}^{d_1 \times d_2}$ to a vector $v_A = \operatorname{vec}(A)$ in $\mathbb{R}^{d_1 d_2}$. Note that we shall adopt the *row-major* convention, and thus the rows of A are stacked together to form v_A. We also recall the standard identity:

$$\operatorname{vec}(E_1 E_2 E_3) = (E_1 \otimes E_3^\top) \operatorname{vec}(E_2), \tag{19}$$

where $\otimes$ denotes the matrix Kronecker product, and E_1, E_2, E_3 are matrices whose dimensions are compatible for the multiplication operation. For any square matrix $A \in \mathbb{R}^{(L+1)\times(L+1)}$, we introduce the notation

$$[M]_{\setminus 0} \in \mathbb{R}^{L \times L} \tag{20}$$

to denote the principal minor of M after removing its first row and column.

Stochastic order notation: In our analysis, we use a concept of high-probability bounds known as *stochastic domination*. This notion, first introduced in [40, 41], provides a convenient way to account for low-probability exceptional events where some bounds may not hold. Consider two families of nonnegative random variables:

$$X = \big(X^{(d)}(u) : d \in \mathbb{N}, u \in U^{(d)}\big), \quad Y = \big(Y^{(d)}(u) : d \in \mathbb{N}, u \in U^{(d)}\big),$$

where $U^{(d)}$ is a possibly d-dependent parameter set. We say that X is *stochastically dominated* by Y, uniformly in u, if for every (small) $\varepsilon > 0$ and (large) $D > 0$ we have

$$\sup_{u \in U^{(d)}} \mathbb{P}[X^{(d)}(u) > d^{\varepsilon} Y^{(d)}(u)] \leq d^{-D}$$

for sufficiently large $d \geq d_0(\varepsilon, D)$. If X is stochastically dominated by Y, uniformly in u, we use the notation $X \prec Y$. Moreover, if for some family X we have $|X| \prec Y$, we also write $X = \mathcal{O}_{\prec}(Y)$.

We also use the notation $X \simeq Y$ to indicate that two families of random variables X, Y are asymptotically equivalent. Precisely, $X \simeq Y$, if there exists $\varepsilon > 0$ such that for every $D > 0$ we have

$$\mathbb{P}\left[|X - Y| > d^{-\varepsilon}\right] \leq d^{-D} \tag{21}$$

for all sufficiently large $d > d_0(\varepsilon, D)$.

SI-10 Moment Calculations and Generalization Errors

For a given set of parameters Γ, its generalization error is defined as

$$e(\Gamma) = \mathbb{E}_{\mathcal{P}_{\text{test}}}\left[\big(y_{\ell+1} - \langle \Gamma, H_Z \rangle\big)^2\right], \tag{22}$$

where $(Z, y_{\ell+1}) \sim \mathcal{P}_{\text{test}}$ is a new sample drawn from the distribution of the test data set. Recall that Z is the input embedding matrix defined in (2) in the main text, and $y_{\ell+1}$ denotes the missing value to be predicted. The goal of this section is to derive an expression for the generalization error $e(\Gamma)$.

Note that the test distribution $\mathcal{P}_{\text{test}}$ crucially depends on the probability distribution of the task vector w used in the linear model in (1). For the ICL test task, we have $w \sim \text{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$, the uniform distribution on the sphere . In what follows, we slightly abuse the notation by writing $w \sim \mathcal{P}_{\text{test}}$ to indicate that w is sampled from the task vector distribution associated with $\mathcal{P}_{\text{test}}$.

Let w be the task vector used in the input matrix Z. Throughout the paper, we use $\mathbb{E}_w[\cdot]$ to denote the conditional expectation with respect to the randomness in the data vectors $\{x_i\}_{i\in[\ell+1]}$ and the noise $\{\epsilon_i\}_{i\in[\ell+1]}$, with the task vector w kept fixed. We have the following expressions for the first two *conditional* moments of $(H_Z, y_{\ell+1})$.

Lemma 1 (Conditional moments)**.** *Let the task vector $w \in$ be fixed. We have*

$$\mathbb{E}_w[y_{\ell+1}] = 0, \quad \textit{and} \quad \mathbb{E}_w[H_Z] = 0. \tag{23}$$

Moreover,

$$\mathbb{E}_w[y_{\ell+1} H_Z] = \frac{1}{d} w \begin{bmatrix} w^\top, & 1+\rho \end{bmatrix} \tag{24}$$

and

$$\mathbb{E}_w\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right] = \frac{(1+\rho)}{d} I_d \otimes \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} w w^\top & (1+2\ell^{-1}) w \\ (1+2\ell^{-1}) w^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix}. \tag{25}$$

Proof. Using the equivalent representations in (162) and (163), it is straightforward to verify the estimates of the first (conditional) moments in (23). To show (24), we note that

$$H_Z = (d/\ell) z_a z_b^\top, \tag{26}$$

where

$$z_a = M_w \begin{bmatrix} s \\ u \end{bmatrix} \qquad \text{and} \qquad z_b = \begin{bmatrix} M_w h \\ (\theta_w a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \end{bmatrix}. \tag{27}$$

Using the representation in (163), we have

$$\mathbb{E}_w \left[y_{\ell+1} H_Z \right] = (d/\ell) \mathbb{E}_w \left[y_{\ell+1} z_a \right] \mathbb{E}_w \left[z_b^\top \right]. \tag{28}$$

Computing the expectations $\mathbb{E}_w \left[y_{\ell+1} z_a \right]$ and $\mathbb{E}_w \left[z_b^\top \right]$ then gives us (24). Next, we show (25). Since z_a and z_b are *independent*,

$$\mathbb{E} \left[\operatorname{vec}(H_Z) \operatorname{vec}(H_Z)^\top \right] = (d/\ell)^2 \, \mathbb{E} \left[z_a z_a^\top \right] \otimes \mathbb{E} \left[z_b z_b^\top \right]. \tag{29}$$

The first expectation on the right-hand side is easy to compute. Since M_w is an orthonormal matrix,

$$\mathbb{E}_w \left[z_a z_a^\top \right] = I_d \tag{30}$$

To obtain the second expectation on the right-hand side of the above expression, we can first verify that

$$\mathbb{E}_w \left[M_w h h^\top M_w \right] = \frac{\ell}{d^2} \left[(1+\rho) I_d + \frac{(\ell+1)}{d} w w^\top \right]. \tag{31}$$

Moreover,

$$\mathbb{E}_w \left[M_w h \left((a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \right) \right] = \frac{\ell(\ell+2)(1+\rho)}{d^3} w \tag{32}$$

and

$$\mathbb{E}_w \left[\left((a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \right)^2 \right] = \frac{\ell(\ell+2)(1+\rho)^2}{d^3}. \tag{33}$$

Combining (31), (32), and (33), we have

$$\mathbb{E} \left[z_b z_b^\top \right] = \frac{(\ell/d)^2 (1+\rho)}{d} \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} w w^\top & (1+2\ell^{-1}) w \\ (1+2\ell^{-1}) w^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix}. \tag{34}$$

Substituting (30) and (34) into (25), we reach the formula in (25). □

Proposition 1 (Generalization error)**.** *For a given weight matrix* Γ*, the generalization error of the linear transformer is*

$$\begin{aligned} e(\Gamma) = \frac{1+\rho}{d} \operatorname{tr} &\left(\Gamma \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} R_{\text{test}} & (1+2\ell^{-1}) b_{\text{test}} \\ (1+2\ell^{-1}) b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix} \Gamma^\top \right) \\ &- \frac{2}{d} \operatorname{tr} \left(\Gamma \begin{bmatrix} R_{\text{test}} \\ (1+\rho) b_{\text{test}}^\top \end{bmatrix} \right) + 1 + \rho, \end{aligned} \tag{35}$$

where

$$b_{\text{test}} := \mathbb{E}_{w \sim \mathcal{P}_{\text{test}}} [w] \quad \textit{and} \quad R_{\text{test}} := \mathbb{E}_{w \sim \mathcal{P}_{\text{test}}} \left[w w^\top \right]. \tag{36}$$

Remark 1. *We use* $w \sim \mathcal{P}_{\text{test}}$ *to indicate that* w *is sampled from the task vector distribution associated with* $\mathcal{P}_{\text{test}}$*. Recall our discussions in Section 2. For the ICL test task, and for the purposes of analytical tractibility, we take* $w \sim$ *Unif*$(\mathcal{S}^{d-1}(\sqrt{d}))$*. In high dimensions, the characterisation of ICL error using* $w \sim$ *Unif*$(\mathcal{S}^{d-1}(\sqrt{d}))$ *will be identical to using* $w \sim \mathcal{N}(0, \mathbb{I})$*. It is then straightforward to check that we want*

$$\textit{(ICL)}: \qquad b_{\text{test}} = 0 \quad \textit{and} \quad R_{\text{test}} = I_d. \tag{37}$$

Proof. Recall the definition of the generalization error in (22). We start by writing

$$e(\Gamma) = \operatorname{vec}(\Gamma)^\top \mathbb{E} \left[\operatorname{vec}(H_Z) \operatorname{vec}(H_Z)^\top \right] \operatorname{vec}(\Gamma) - 2 \operatorname{vec}(\Gamma)^\top \operatorname{vec}(\mathbb{E} \left[y_{N+1} H_Z \right]) + \mathbb{E} \left[y_{\ell+1}^2 \right], \tag{38}$$

where H_Z is a matrix in the form of (9) and H_Z is independent of Γ. Since $y_{\ell+1} = x_{\ell+1}^\top w + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \rho)$ denoting the noise, it is straightforward to check that

$$\mathbb{E}\left[y_{\ell+1}^2\right] = 1 + \rho. \tag{39}$$

Using the moment estimate (25) in Lemma 1 and the identity (19), we have

$$\begin{aligned}&\operatorname{vec}(\Gamma)^\top \mathbb{E}\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right]\operatorname{vec}(\Gamma)\\ &\quad= \frac{1+\rho}{d}\operatorname{tr}\left(\Gamma\begin{bmatrix}\frac{d}{\ell}I_d + (1+\ell^{-1})(1+\rho)^{-1}R_{\text{test}} & (1+2\ell^{-1})b_{\text{test}}\\ (1+2\ell^{-1})b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho)\end{bmatrix}\Gamma^\top\right).\end{aligned} \tag{40}$$

Moreover, by (24),

$$\operatorname{vec}(\Gamma)^\top \operatorname{vec}\left(\mathbb{E}\left[y_{\ell+1}H_Z\right]\right) = \frac{1}{d}\operatorname{tr}\left(\Gamma\begin{bmatrix}R_{\text{test}}\\ (1+\rho)b_{\text{test}}^\top\end{bmatrix}\right). \tag{41}$$

□

Corollary 1. *For a given set of parameters Γ, its generalization error can be written as*

$$e(\Gamma) = \frac{1}{d}\operatorname{tr}\left(\Gamma B_{\text{test}}\Gamma^\top\right) - \frac{2}{d}\operatorname{tr}\left(\Gamma A_{\text{test}}^\top\right) + (1+\rho) + \mathcal{E}, \tag{42}$$

where

$$A_{\text{test}} := \begin{bmatrix}R_{\text{test}} & (1+\rho)b_{\text{test}}\end{bmatrix}, \tag{43}$$

$$B_{\text{test}} := \begin{bmatrix}\frac{1}{\alpha}(1+\rho)I_d + R_{\text{test}} & (1+\rho)b_{\text{test}}\\ (1+\rho)b_{\text{test}}^\top & (1+\rho)^2\end{bmatrix}, \tag{44}$$

and $R_{\text{test}}, b_{\text{test}}$ are as defined in (36). Moreover, $\mathcal{E}$ denotes an "error" term such that

$$|\mathcal{E}| \leq \frac{C_{\alpha,\rho}\max\left\{\left\|R_{\text{test}}\right\|_{\mathsf{op}}, \left\|b_{\text{test}}\right\|, 1\right\}\left(\left\|\Gamma\right\|_{\mathsf{F}}^2/d\right)}{d}, \tag{45}$$

where $C_{\alpha,\rho}$ is some constant that only depends on α and ρ.

Proof. Let

$$\Delta = \begin{bmatrix}\frac{d}{\ell}(1+\rho)I_d + (1+\ell^{-1})R_{\text{test}} & (1+2\ell^{-1})(1+\rho)b_{\text{test}}\\ (1+2\ell^{-1})(1+\rho)b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho)^2\end{bmatrix} - B_{\text{test}}. \tag{46}$$

It is straightforward to check that

$$\begin{aligned}\mathcal{E} &= \frac{1}{d}\operatorname{tr}\left(\Gamma\Delta\Gamma^\top\right) &(47)\\ &= \frac{1}{d}\operatorname{vec}(\Gamma)^\top(I_d \otimes \Delta)\operatorname{vec}(\Gamma) &(48)\\ &\leq \left\|\Delta\right\|_{\mathsf{op}}\frac{\left\|\Gamma\right\|_{\mathsf{F}}^2}{d}. &(49)\end{aligned}$$

The bound in (45) follows from the estimate that $\left\|\Delta\right\|_{\mathsf{op}} \leq C_{\alpha,\rho}\max\left\{\left\|R_{\text{test}}\right\|_{\mathsf{op}}, \left\|b_{\text{test}}\right\|, 1\right\}/d$. □

Remark 2. *Consider the optimal weight matrix Γ^* obtained by solving the ridge regression problem in (10). Since Γ^* is the optimal solution of (10), we must have*

$$\frac{n}{d}\lambda\left\|\Gamma^*\right\|_{\mathsf{F}}^2 \leq \sum_{\mu\in[n]}(y_{\ell+1}^\mu)^2, \tag{50}$$

where the right-hand side is the value of the objective function of (10) when we choose Γ to be the all-zero matrix. It follows that

$$\frac{\left\|\Gamma^*\right\|_{\mathsf{F}}^2}{d} \leq \frac{\sum_{\mu\in[n]}(y_{\ell+1}^\mu)^2}{\lambda n}. \tag{51}$$

By the law of large numbers, $\frac{\sum_{\mu\in[n]}y_\mu^2}{n} \to 1+\rho$ as $n \to \infty$. Thus, $\left\|\Gamma^*\right\|_{\mathsf{F}}^2/d$ is asymptotically bounded by the constant $(1+\rho)/\lambda$. Furthermore, it is easy to check that $\left\|R_{\text{test}}\right\|_{\mathsf{op}} = \mathcal{O}(1)$ and $\left\|b_{\text{test}}\right\| = \mathcal{O}(1)$ for the ICL task [see (37)]. It then follows from Corollary 1 that the generalization error associated with the optimal parameters Γ^* is asymptotically determined by the first three terms on the right-hand side of (42).

SI-11 Analysis of Ridge Regression: Extended Resolvent Matrices

We see from Corollary 1 and Remark 2 that the two key quantities in determining the generalization error $e(\Gamma^*)$ are

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\text{test}}^\top) \qquad \text{and} \qquad \frac{1}{d}\operatorname{tr}(\Gamma^* B_{\text{test}}(\Gamma^*)^\top), \tag{52}$$

where A_{test} and B_{test} are the matrices defined in (43) and (44), respectively. In this section, we show that the two quantities in (52) can be obtained by studying a parameterized family of extended resolvent matrices.

To start, we observe that the ridge regression problem in (7) admits the following closed-form solution:

$$\operatorname{vec}(\Gamma^*) = G\left(\textstyle\sum_{\mu\in[n]} y_\mu \operatorname{vec}(H_\mu)\right)/d, \tag{53}$$

where G is a resolvent matrix defined as

$$G = \left(\textstyle\sum_{\mu\in[n]} \operatorname{vec}(H_\mu)\operatorname{vec}(H_\mu)^\top/d + \tau\lambda I\right)^{-1}. \tag{54}$$

For our later analysis of the generalization error, we need to consider a more general, "parameterized" version of G, defined as

$$G(\pi) = \left(\textstyle\sum_{\mu\in[n]} \operatorname{vec}(H_\mu)\operatorname{vec}(H_\mu)^\top/d + \pi\Pi + \tau\lambda I\right)^{-1}, \tag{55}$$

where $\Pi \in \mathbb{R}^{(d^2+d)\times(d^2+d)}$ is a symmetric positive-semidefinite matrix and π is a nonnegative scalar. The original resolvent G in (54) is a special case, corresponding to $\pi = 0$.

The objects in (53) and (55) are the submatrices of an *extended* resolvent matrix, which we construct as follows. For each $\mu \in [n]$, let

$$z_\mu = \begin{bmatrix} y_\mu/d \\ \operatorname{vec}(H_\mu)/\sqrt{d} \end{bmatrix} \tag{56}$$

be an (d^2+d+1)-dimensional vector. Let

$$\Pi_e = \begin{bmatrix} 0 & \\ & \Pi \end{bmatrix}, \tag{57}$$

where Π is the $(d^2+d)\times(d^2+d)$ matrix in (55). Define an extended resolvent matrix

$$G_e(\pi) = \frac{1}{\sum_{\mu\in[n]} z_\mu z_\mu^\top + \pi\Pi_e + \tau\lambda I}. \tag{58}$$

By block-matrix inversion, it is straightforward to check that

$$G_e(\pi) = \begin{bmatrix} c(\pi) & -c(\pi)q^\top(\pi) \\ -c(\pi)q(\pi) & G(\pi) + c(\pi)q(\pi)q^\top(\pi) \end{bmatrix}, \tag{59}$$

where

$$q(\pi) := \frac{1}{d^{3/2}} G(\pi)\left(\textstyle\sum_{\mu\in[n]} y_\mu \operatorname{vec}(H_\mu)\right) \tag{60}$$

is a vector in $\mathbb{R}^{d(d+1)}$, and $c(\pi)$ is a scalar such that

$$\frac{1}{c(\pi)} = \frac{1}{d^2}\sum_{\mu\in[n]} y_\mu^2 + \tau\lambda - \frac{1}{d^3}\sum_{\mu,\nu\in[n]} y_\mu y_\nu \operatorname{vec}(H_\mu)^\top G(\pi)\operatorname{vec}(H_\nu). \tag{61}$$

By comparing (60) with (53), we see that

$$\operatorname{vec}(\Gamma^*) = \sqrt{d}\, q(0). \tag{62}$$

Moreover, as shown in the following lemma, the two key quantities in (52) can also be obtained from the extended resolvent $G_e(\pi)$.

Lemma 2. *For any matrix $A \in \mathbb{R}^{d\times(d+1)}$,*

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A^\top) = \frac{-1}{c(0)\sqrt{d}}\begin{bmatrix}0 & \operatorname{vec}(A)^T\end{bmatrix} G_e(0)e_1, \tag{63}$$

where e_1 denotes the first natural basis vector in $\mathbb{R}^{d^2+d+1}$. Moreover, for any symmetric and positive semidefinite matrix $B \in \mathbb{R}^{(d+1)\times(d+1)}$, if we set

$$\Pi = I_d \otimes B \tag{64}$$

in (57), then

$$\frac{1}{d}\operatorname{tr}(\Gamma^* B (\Gamma^*)^\top) = \frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c(\pi)}\right)\Big|_{\pi=0}. \tag{65}$$

Proof. The identity (63) follows immediately from the block form of $G_e(\pi)$ in (59) and the observation in (62). To show (65), we take the derivative of $1/c(\pi)$ with respect to π. From (61), and using the identity

$$\frac{\mathrm{d}}{\mathrm{d}\pi}G(\pi) = -G(\pi)\Pi G(\pi), \tag{66}$$

we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c(\pi)}\right) &= \frac{1}{d^3}\sum_{\mu,\nu\in[n]} y_\mu y_\nu \operatorname{vec}(H_\mu)^\top G(\pi)\Pi G(\pi)\operatorname{vec}(H_\nu) &(67)\\ &= q^\top(\pi)\Pi q(\pi). &(68)\end{aligned}$$

Thus, by (62),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c(\pi)}\right)\Big|_{\pi=0} &= \frac{1}{d}\left(\operatorname{vec}(\Gamma^*)\right)^\top \Pi \operatorname{vec}(\Gamma^*) &(69)\\ &= \frac{1}{d}\left(\operatorname{vec}(\Gamma^*)\right)^\top (I_d\otimes B)\operatorname{vec}(\Gamma^*). &(70)\end{aligned}$$

Applying the identity in (19) to the right-hand side of the above equation, we reach (65). □

Remark 3. *To lighten the notation, we will often write $G_e(\pi)$ [resp. $G(\pi)$] as G_e [resp. G], leaving their dependence on the parameter π implicit.*

Remark 4. *In light of (64) and (65), we will always choose*

$$\Pi = I_d \otimes B_{\text{test}}, \tag{71}$$

where B_{test} is the matrix defined in (44).

SI-12 An Asymptotic Equivalent of the Extended Resolvent Matrix

In this section, we derive an asymptotic equivalent of the extended resolvent G_e defined in (58). From this equivalent version, we can then obtain the asymptotic limits of the right-hand sides of (63) and (65). Our analysis relies on non-rigorous but technically sound heuristic arguments from random matrix theory. Therefore, we refer to our theoretical predictions as *results* rather than propositions.

Recall that there are k unique task vectors $\{w_i\}_{i\in[k]}$ in the training set. Let

$$b_{\text{tr}} := \frac{1}{k}\sum_{i\in[k]} w_i \quad \text{and} \quad R_{\text{tr}} := \frac{1}{k}\sum_{i\in[k]} w_i w_i^\top \tag{72}$$

denote the empirical mean and correlation matrix of these k regression vectors, respectively. Define

$$A_{\text{tr}} := \begin{bmatrix} R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}. \tag{73}$$

and

$$E_{\text{tr}} := \begin{bmatrix} \frac{(1+\rho)}{\alpha}I_d + R_{\text{tr}} & (1+\rho)b_{\text{tr}} \\ (1+\rho)b_{\text{tr}}^\top & (1+\rho)^2\end{bmatrix}. \tag{74}$$

Definition 1. *Consider the extended resolvent $G_e(\pi)$ in (58), with Π_e chosen in the forms of (57) and (71). Let $\widetilde{G}_e$ be another matrix of the same size as $G_e(\pi)$. We say that $\widetilde{G}_e$ and $G_e(\pi)$ are asymptotically equivalent, if the following conditions hold.*

(1) For any two deterministic and unit-norm vectors $u, v \in \mathbb{R}^{d^2+d+1}$,

$$u^\top G_e(\pi) v \simeq u^\top \widetilde{G}_e v, \tag{75}$$

where $\simeq$ is the asymptotic equivalent notation defined in (21).

(2) Let $A_{\mathrm{tr}} = \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}$. For any deterministic, unit-norm vector $v \in \mathbb{R}^{d^2+d+1}$,

$$\frac{1}{\sqrt{d}} \begin{bmatrix} 0 & \mathrm{vec}(A_{\mathrm{tr}})^\top \end{bmatrix} G_e(\pi) v \simeq \frac{1}{\sqrt{d}} \begin{bmatrix} 0 & \mathrm{vec}(A_{\mathrm{tr}})^\top \end{bmatrix} \widetilde{G}_e v. \tag{76}$$

(3) Recall the notation introduced in (20). We have

$$\frac{1}{d^2} \operatorname{tr}\left(\big[G_e(\pi)\big]_{\backslash 0} \cdot [I \otimes E_{\mathrm{tr}}] \right) = \frac{1}{d^2} \operatorname{tr}\left(\Big[\widetilde{G}_e\Big]_{\backslash 0} \cdot [I \otimes E_{\mathrm{tr}}] \right) + \mathcal{O}_\prec(d^{-1/2}), \tag{77}$$

where $\big[G_e(\pi)\big]_{\backslash 0}$ and $\big[\mathcal{G}_e(\pi)\big]_{\backslash 0}$ denote the principal minors of $G_e(\pi)$ and $\mathcal{G}_e(\pi)$, respectively.

Result 2. *Let χ_π denote the unique positive solution to the equation*

$$\chi_\pi = \frac{1}{d} \operatorname{tr}\left[\left(\frac{\tau}{1+\chi_\pi} E_{\mathrm{tr}} + \pi B_{\mathrm{test}} + \lambda \tau I_d \right)^{-1} E_{\mathrm{tr}} \right], \tag{78}$$

where B_{test} is the positive-semidefinite matrix in (44), with $b_{\mathrm{test}}, R_{\mathrm{test}}$ chosen according to (37). The extended resolvent $G_e(\pi)$ in (58) is asymptotically equivalent to

$$\mathcal{G}_e(\pi) := \left(\frac{\tau}{1+\chi_\pi} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} \mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} \right)^\top \\ \frac{1}{\sqrt{d}} \mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} \right) & I_d \otimes E_{\mathrm{tr}} \end{bmatrix} + \pi \Pi_e + \tau \lambda I \right)^{-1}, \tag{79}$$

in the sense of Definition 1. In the above expression, Π_e is the matrix in (57) with $\Pi = I_d \otimes B_{\mathrm{test}}$.

In what follows, we present the steps in reaching the asymptotic equivalent $\mathcal{G}_e(\pi)$ given in (79). To start, let $G_e^{[\mu]}$ to denote a "leave-one-out" version of G_e, defined as

$$G_e^{[\mu]} = \frac{1}{\sum_{\nu \neq \mu} z_\nu z_\nu^\top + \pi \Pi_e + \tau \lambda I}. \tag{80}$$

By (58), we have

$$G_e \left(\textstyle\sum_{\mu \in [n]} z_\mu z_\mu^\top + \pi \Pi_e + \tau \lambda I \right) = I. \tag{81}$$

Applying the Woodbury matrix identity then gives us

$$\sum_{\mu \in [n]} \frac{1}{1 + z_\mu^\top G_e^{[\mu]} z_\mu} G_e^{[\mu]} z_\mu z_\mu^\top + G_e(\pi \Pi_e + \tau \lambda I) = I. \tag{82}$$

To proceed, we study the quadratic form $z_\mu^\top G_e^{[\mu]} z_\mu$. Let w_μ denotes the task vector associated with z_μ. Conditioned on w_μ and G_e^μ, the quadratic form $z_\mu^\top G_e^{[\mu]} z_\mu$ concentrates around its *conditional expectation* with respect to the remaining randomness in z_μ. Specifically,

$$z_\mu^\top G_e^{[\mu]} z_\mu = \chi^\mu(w_\mu) + \mathcal{O}_\prec(d^{-1/2}), \tag{83}$$

where

$$\chi^\mu(w_\mu) := \frac{1}{d^2} \operatorname{tr}\left([G_e^\mu]_{\backslash 0} \cdot \big[I \otimes E(w_\mu)\big] \right), \tag{84}$$

and

$$E(w) \coloneqq \begin{bmatrix} \frac{1+\rho}{\alpha} I_d + ww^\top & (1+\rho)w \\ (1+\rho)w^\top & (1+\rho)^2 \end{bmatrix}. \tag{85}$$

Substituting $z_\mu^\top G_e^{[\mu]} z_\mu$ in (82) by $\chi^\mu(w_\mu)$, we get

$$\sum_{\mu \in [n]} \frac{1}{1+\chi^\mu(w_\mu)} G_e^{[\mu]} z_\mu z_\mu^\top + G_e(\pi \Pi_e + \tau\lambda I) = I + \Delta_1, \tag{86}$$

where

$$\Delta_1 \coloneqq \sum_{\mu \in [n]} \frac{z_\mu^\top G_e^{[\mu]} z_\mu - \chi^\mu(w_\mu)}{(1+\chi^\mu(w_\mu))(1+z_\mu^\top G_e^{[\mu]} z_\mu)} G_e^{[\mu]} z_\mu z_\mu^\top \tag{87}$$

is a matrix that captures the approximation error of the above substitution.

Next, we replace $z_\mu z_\mu^\top$ on the left-hand side of (86) by its *conditional* expectation $\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right]$, conditioned on the task vector w_μ. This allows us to rewrite (86) as

$$\sum_{\mu \in [n]} \frac{1}{1+\chi^\mu(w_\mu)} G_e^{[\mu]} \mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] + G_e(\pi \Pi_e + \tau\lambda I) = I + \Delta_1 + \Delta_2, \tag{88}$$

where

$$\Delta_2 \coloneqq \sum_{\mu \in [n]} \frac{1}{1+\chi^\mu(w_\mu)} G_e^{[\mu]} \left(\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] - z_\mu z_\mu^\top\right) \tag{89}$$

captures the corresponding approximation error. Recall the definition of z_μ in (56). Using the moment estimates in Lemma 1, we have

$$\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] = \frac{1}{d^2} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + \frac{1}{d^2} \begin{bmatrix} 0 & \\ & I_d \otimes \mathcal{E}_\mu \end{bmatrix}, \tag{90}$$

where $E(w_\mu)$ is the matrix defined in (85) and

$$\mathcal{E}_\mu = \frac{1}{\ell} \begin{bmatrix} w_\mu w_\mu^\top & 2(1+\rho)w_\mu \\ 2(1+\rho)w_\mu^\top & 2(1+\rho)^2 \end{bmatrix}. \tag{91}$$

Replacing the conditional expectation $\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right]$ in (88) by the main (i.e. the first) term on the right-hand side of (90), we can transform (88) to

$$\frac{\tau}{n} \sum_{\mu \in [n]} \frac{1}{1+\chi^\mu(w_\mu)} G_e^{[\mu]} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + G_e(\pi \Pi_e + \tau\lambda I) = I + \Delta_1 + \Delta_2 + \Delta_3, \tag{92}$$

where we recall $\tau = n/d^2$, and we use Δ_3 to capture the approximation error associated with $\mathcal{E}_\mu$.

Next, we replace the "leave-one-out" terms G_e^μ and $\chi^\mu(w_\mu)$ in (92) by their "full" versions. Specifically, we replace G_e^μ by G_e, and $\chi^\mu(w_\mu)$ by

$$\chi(w_\mu) \coloneqq \frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot \left[I \otimes E(w_\mu)\right]\right). \tag{93}$$

It is important to note the difference between (84) and (93): the former uses G_e^μ and the latter G_e. After these replacements and using Δ_4 to capture the approximation errors, we have

$$G_e \left(\frac{\tau}{n} \sum_{\mu \in [n]} \frac{1}{1+\chi(w_\mu)} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + \pi \Pi_e + \tau\lambda I \right) = I + \sum_{j \leq 4} \Delta_j. \tag{94}$$

Recall that there are k unique task vectors $\{w_i\}_{1\leq i\leq k}$ in the training set consisting of n input samples. Each sample is associated with one of these task vectors, sampled uniformly from the set $\{w_i\}_{1\leq i\leq k}$. In our analysis, we shall assume that k divides n and that each unique task vector is associated with exactly n/k input samples. (We note that this assumption merely serves to simplify the notation. The asymptotic characterization of the random matrix G_e remains the same even without this assumption.) Observe that there are only k unique terms in the sum on the left-hand side of (94). Thus,

$$G_e\left(\frac{\tau}{k}\sum_{i\in[k]}\frac{1}{1+\chi(w_i)}\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}w_i^\top\otimes\begin{bmatrix}w_i^\top & 1+\rho\end{bmatrix}\\ \frac{1}{\sqrt{d}}w_i\otimes\begin{bmatrix}w_i\\ 1+\rho\end{bmatrix} & I_d\otimes E(w_i)\end{bmatrix}+\pi\Pi_e+\tau\lambda I\right)=I+\sum_{j\leq 4}\Delta_j. \tag{95}$$

So far, we have been treating the k task vectors $\{w_i\}_{i\in[k]}$ as fixed vectors, only using the randomness in the input samples that are associated with the data vectors $\{x_i^\mu\}$. To further simplify our asymptotic characterization, we take advantage of the fact that $\{w_i\}_{i\in[k]}$ are independently sampled from $\mathrm{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$. To that end, we can first show that $\chi(w_i)$ in (93) concentrates around its expectation. Specifically,

$$\chi(w_i)=\mathbb{E}\left[\frac{1}{d^2}\operatorname{tr}\left([G_e]_{\setminus 0}\cdot\left[I\otimes E(w_i)\right]\right)\right]+\mathcal{O}_\prec(d^{-1/2}). \tag{96}$$

By symmetry, we must have

$$\mathbb{E}\left[\frac{1}{d^2}\operatorname{tr}\left([G_e]_{\setminus 0}\cdot\left[I\otimes E(w_i)\right]\right)\right]=\mathbb{E}\left[\frac{1}{d^2}\operatorname{tr}\left([G_e]_{\setminus 0}\cdot\left[I\otimes E(w_j)\right]\right)\right] \tag{97}$$

for any $1\leq i<j\leq k$. It follows that $\left|\chi(w_i)-\chi(w_j)\right|=\mathcal{O}_\prec(d^{-1/2})$, and thus, by a union bound,

$$\max_{i\in[k]}\left|\chi(w_{k_1})-\widehat{\chi}_{\text{ave}}\right|=\mathcal{O}_\prec(d^{-1/2}), \tag{98}$$

where

$$\widehat{\chi}_{\text{ave}}:=\frac{1}{k}\sum_{i\in[k]}\chi(w_i). \tag{99}$$

Upon substituting (93) into (99), it is straightforward to verify the following characterization of $\widehat{\chi}_{\text{ave}}$:

$$\widehat{\chi}_{\text{ave}}=\frac{1}{d^2}\operatorname{tr}\left([G_e]_{\setminus 0}\cdot[I\otimes E_{\text{tr}}]\right). \tag{100}$$

The estimate in (98) prompts us to replace the terms $\chi(w_i)$ in the right-hand side of (95) by the common value $\widehat{\chi}_{\text{ave}}$. As before, we introduce a matrix Δ_5 to capture the approximation error associated with this step. Using the newly introduced notation $E_{\text{tr}}, b_{\text{tr}}$ and R_{tr} in (74) and (72), we can then simplify (95) as

$$G_e\left(\frac{\tau}{1+\widehat{\chi}_{\text{ave}}}\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}\right)^\top\\ \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}\right) & I_d\otimes E_{\text{tr}}\end{bmatrix}+\pi\Pi_e+\tau\lambda I\right)$$
$$=I+\sum_{1\leq j\leq 5}\Delta_j. \tag{101}$$

Define

$$\widehat{\mathcal{G}}_e(\pi):=\left(\frac{\tau}{1+\widehat{\chi}_{\text{ave}}}\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}\right)^\top\\ \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}\right) & I_d\otimes E_{\text{tr}}\end{bmatrix}+\pi\Pi_e+\tau\lambda I\right)^{-1}. \tag{102}$$

Then

$$G_e=\widehat{\mathcal{G}}_e(\pi)+\underbrace{\widehat{\mathcal{G}}_e(\pi)\left(\Delta_1+\Delta_2+\Delta_3+\Delta_4+\Delta_5\right)}_{\text{approximation errors}}. \tag{103}$$

Remark 5. *We claim that $\widehat{\mathcal{G}}_e$ is asymptotically equivalent to G_e, in the sense of Definition 1. Given (103), proving this claim requires showing that, for $j = 1, 2, \ldots, 5$,*

$$u^\top \left(\widehat{\mathcal{G}}_e(\pi) \Delta_j \right) v \simeq 0, \tag{104a}$$

$$\frac{1}{\sqrt{d}} \begin{bmatrix} 0 & \mathrm{vec}(A_{\mathrm{tr}})^\top \end{bmatrix} \left(\widehat{\mathcal{G}}_e(\pi) \Delta_j \right) v \simeq 0, \tag{104b}$$

and

$$\frac{1}{d^2} \operatorname{tr} \left(\left[\widehat{\mathcal{G}}_e(\pi) \Delta_j \right]_{\setminus 0} \cdot [I \otimes E_{\mathrm{tr}}] \right) \simeq 0, \tag{104c}$$

for any deterministic and unit-norm vectors u, v and for $A_{\mathrm{tr}} = \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}$.

We note the equivalent matrix $\widehat{\mathcal{G}}_e(\pi)$ still involves one scalar $\widehat{\chi}_{\mathrm{ave}}$ that depends on the original resolvent $G_e(\pi)$. Next, we show that $\widehat{\chi}_{\mathrm{ave}}$ can be replaced by χ_π, the unique positive solution to (78). To that end, we recall the characterization in (100). Using the claim that $G_e(\pi)$ and $\widehat{\mathcal{G}}_e(\pi)$ are asymptotically equivalent (in particular, in the sense of (77)), we have

$$\widehat{\chi}_{\mathrm{ave}} \simeq \frac{1}{d^2} \operatorname{tr} \left(\left[\widehat{\mathcal{G}}_e(\pi) \right]_{\setminus 0} \cdot [I \otimes E_{\mathrm{tr}}] \right). \tag{105}$$

To compute the first term on the right-hand side of the above estimate, we directly invert the block matrix $\widehat{\mathcal{G}}_e(\pi)$ in (102). Recall that Π_e is chosen in the forms of (57) and (64). It is then straightforward to verify that

$$\widehat{\mathcal{G}}_e = \begin{bmatrix} \bar{c} & -\bar{c}\,\bar{q}^\top \\ -\bar{c}\,\bar{q} & I \otimes F_E(\widehat{\chi}_{\mathrm{ave}}) + \bar{c}\,\bar{q}\bar{q}^\top \end{bmatrix}, \tag{106}$$

where $F_E(\chi)$ is a matrix valued function such that

$$F_E(\chi) = \left(\frac{\tau}{1+\chi} E_{\mathrm{tr}} + \pi B + \lambda \tau I_{d+1} \right)^{-1}, \tag{107}$$

$$\bar{q} = \frac{\tau}{(1+\widehat{\chi}_{\mathrm{ave}})\sqrt{d}} \mathrm{vec} \left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} F_E(\widehat{\chi}_{\mathrm{ave}}) \right), \tag{108}$$

and

$$1/\bar{c} = \frac{\tau(1+\rho)}{1+\widehat{\chi}_{\mathrm{ave}}} + \lambda\tau - \frac{\tau^2}{(1+\widehat{\chi}_{\mathrm{ave}})^2 d} \operatorname{tr} \left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} F_E(\widehat{\chi}_{\mathrm{ave}}) \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}^\top \right). \tag{109}$$

Using (106), we can now write the equation (105) as

$$\begin{aligned} \widehat{\chi}_{\mathrm{ave}} &\simeq \frac{1}{d} \operatorname{tr} \left(F_E(\widehat{\chi}_{\mathrm{ave}}) E_{\mathrm{tr}} \right) \\ &\quad + \frac{\bar{c}\tau^2}{(1+\widehat{\chi}_{\mathrm{ave}})^2 d^3} \operatorname{tr} \left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} F_E(\widehat{\chi}_{\mathrm{ave}}) E_{\mathrm{tr}} F_E(\widehat{\chi}_{\mathrm{ave}}) \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}^\top \right). \end{aligned} \tag{110}$$

The second term on the right-hand side of (110) is negligible. Indeed,

$$\begin{aligned} &\operatorname{tr} \left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} F_E(\widehat{\chi}_{\mathrm{ave}}) E_{\mathrm{tr}} F_E(\widehat{\chi}_{\mathrm{ave}}) \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}^\top \right) \\ &\qquad \leq \left\| F_E(\widehat{\chi}_{\mathrm{ave}}) E_{\mathrm{tr}} F_E(\widehat{\chi}_{\mathrm{ave}}) \right\|_{\mathsf{op}} (\|R_{\mathrm{tr}}\|_{\mathsf{F}}^2 + (1+\rho)^2 \|b_{\mathrm{tr}}\|^2). \end{aligned} \tag{111}$$

By construction, $\left\| F_E(\widehat{\chi}_{\mathrm{ave}}) \right\|_{\mathsf{op}} \leq (\lambda\tau)^{-1}$. Moreover, since the task vectors $\{w_i\}_{i \in [k]}$ are independent vectors sampled from $\mathrm{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$, it is easy to verify that

$$\|E_{\mathrm{tr}}\|_{\mathsf{op}} = \mathcal{O}_\prec(1), \qquad \|R_{\mathrm{tr}}\|_{\mathsf{F}} = \mathcal{O}_\prec(\sqrt{d}) \qquad \text{and} \qquad \|b_{\mathrm{tr}}\|_2 = \mathcal{O}_\prec(1). \tag{112}$$

Finally, since $\bar{c}$ is an element of $\widehat{\mathcal{G}}_e$, we must have $|\bar{c}| \leq \left\| \widehat{\mathcal{G}}_e \right\|_{\mathsf{op}} \leq (\tau\lambda)^{-1}$. Combining these estimates gives us

$$\frac{\bar{c}\tau^2}{(1+\widehat{\chi}_{\mathrm{ave}})^2 d^3} \operatorname{tr} \left(\begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix} F_E(\widehat{\chi}_{\mathrm{ave}}) E_{\mathrm{tr}} F_E(\widehat{\chi}_{\mathrm{ave}}) \begin{bmatrix} R_{\mathrm{tr}} & (1+\rho) b_{\mathrm{tr}} \end{bmatrix}^\top \right) = \mathcal{O}_\prec(d^{-2}), \tag{113}$$

and thus we can simplify (110) as

$$\widehat{\chi}_{\text{ave}} \simeq \frac{1}{d} \operatorname{tr} \left[\left(\frac{\tau}{1+\widehat{\chi}_{\text{ave}}} E_{\text{tr}} + \pi B + \lambda \tau I_d \right)^{-1} E_{\text{tr}} \right]. \tag{114}$$

Observe that (114) is a small perturbation of the self-consistent equation in (78). By the stability of the equation (78), we then have

$$\widehat{\chi}_{\text{ave}} \simeq \chi_\pi, \tag{115}$$

where χ_π is the unique positive solution to (78).

Recall the definitions of $\mathcal{G}_e(\pi)$ and $\widehat{\mathcal{G}}_e(\pi)$ in (102) and (79), respectively. By the standard resolvent identity,

$$\begin{aligned} &\widehat{\mathcal{G}}_e(\pi) - \mathcal{G}_e(\pi) \\ &= \frac{\tau[\widehat{\chi}_{\text{ave}} - \chi_\pi]}{[1+\chi_\pi][1+\widehat{\chi}_{\text{ave}}]} \widehat{\mathcal{G}}_e(\pi) \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} \operatorname{vec}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} \right)^\top \\ \frac{1}{\sqrt{d}} \operatorname{vec}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} \right) & I_d \otimes E_{\text{tr}} \end{bmatrix} \mathcal{G}_e(\pi). \end{aligned} \tag{116}$$

By construction, $\left\| \widehat{\mathcal{G}}_e(\pi) \right\|_{\mathsf{op}} \le 1/(\tau\lambda)$ and $\left\| \mathcal{G}_e(\pi) \right\|_{\mathsf{op}} \le 1/(\tau\lambda)$. Moreover, $\left\| E_{\text{tr}} \right\|_{\mathsf{op}} \prec 1$ and

$$\left\| \frac{1}{\sqrt{d}} \operatorname{vec}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} \right) \right\| \prec 1. \tag{117}$$

It then follows from (115) and (116) that

$$\left\| \widehat{\mathcal{G}}_e(\pi) - \mathcal{G}_e(\pi) \right\|_{\mathsf{op}} \simeq 0. \tag{118}$$

If $\widehat{\mathcal{G}}_e(\pi)$ satisfies the equivalent conditions (75), (76) and (77) (as claimed in our analysis above), then the estimate in (118) allows us to easily check that $\mathcal{G}_e(\pi)$ also satisfies (75), (76) and (77). Thus, we claim that $\mathcal{G}_e(\pi)$ is asymptotically equivalent to the extended resolvent matrix $G_e(\pi)$ in the sense of Definition 1.

SI-13 Asymptotic Limits of the Generalization Errors

In this section, we use the characterization in Result 2 to derive the asymptotic limits of the generalization errors of associated with the set of parameters Γ^* learned from ridge regression.

SI-13.1 Asymptotic Limits of the Linear and Quadratic Terms

From Corollary 1 and the discussions in Remark 2, characterizing the test error $e(\Gamma^*)$ boils down to computing the linear term $\frac{1}{d} \operatorname{tr} \left(\Gamma^* A_{\text{test}}^\top \right)$ and the quadratic term $\frac{1}{d} \operatorname{tr} \left(\Gamma^* B_{\text{test}} (\Gamma^*)^\top \right)$, where A_{test} and B_{test} are the matrices defined in (43) and (44), respectively.

We consider test data distributions $\mathcal{P}_{\text{test}}$ as follows. From (37), the ICL task test setting we consider corresponds to choosing

$$\text{(ICL)}: \qquad A_{\text{test}} = \begin{bmatrix} I_d & 0 \end{bmatrix} \quad \text{and} \quad B_{\text{test}} = \begin{bmatrix} (\frac{1+\rho}{\alpha} + 1) I_d & \\ & (1+\rho)^2 \end{bmatrix}. \tag{119}$$

.

Result 3. *Let Γ^* be the set of parameters learned from the ridge regression problem in (10). Let $A_{\text{test}} \in \mathbb{R}^{d \times (d+1)}$ and $B_{\text{test}} \in \mathbb{R}^{(d+1)\times(d+1)}$ be two matrices constructed as in (119). We have*

$$\frac{1}{d} \operatorname{tr}(\Gamma^* A_{\text{test}}^\top) \simeq \frac{1}{d} \operatorname{tr}\left(\Gamma_{\text{eq}}^* A_{\text{test}}^\top \right), \tag{120}$$

and

$$\frac{1}{d} \operatorname{tr}(\Gamma^* B_{\text{test}} (\Gamma^*)^\top) \simeq \frac{1}{d} \operatorname{tr}(\Gamma_{\text{eq}}^* B_{\text{test}} (\Gamma_{\text{eq}}^*)^T) - \frac{c_e}{d} \operatorname{tr}\left(B_{\text{test}} \left[(E_{\text{tr}} + \xi I)^{-1} - \xi (E_{\text{tr}} + \xi I)^{-2} \right] \right). \tag{121}$$

In the above displays, Γ^*_{eq} is an asymptotic equivalent of Γ^*, defined as

$$\Gamma^*_{\text{eq}} := \begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} (E_{\text{tr}} + \xi I)^{-1}, \tag{122}$$

where ξ is the unique positive solution to the self-consistent equation

$$\xi \mathcal{M}_\kappa \left(\frac{1+\rho}{\alpha} + \xi \right) - \frac{\tau\lambda}{\xi} = 1 - \tau, \tag{123}$$

and $\mathcal{M}_\kappa(\cdot)$ is the function defined in (181). Moreover, the scalar c_e in (121) is defined as

$$c_e = \frac{\rho + \nu - \nu^2 \mathcal{M}_\kappa(\nu) - \xi \left[1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2 \mathcal{M}'_\kappa(\nu) \right]}{1 - 2\xi\mathcal{M}_\kappa(\nu) - \xi^2 \mathcal{M}'_\kappa(\nu) - \tau}, \tag{124}$$

where

$$\nu := \frac{1+\rho}{\alpha} + \xi. \tag{125}$$

To derive the asymptotic characterizations (120) and (121) in Result 3, we first use block-matrix inversion to rewrite $\mathcal{G}_e(\pi)$ in (79) as

$$\mathcal{G}_e(\pi) = \begin{bmatrix} c^*(\pi) & -c^*(\pi)\,(q^*(\pi))^\top \\ -c^*(\pi)\,q^*(\pi) & I \otimes F_E(\chi_\pi) + c^*(\pi) q^*(\pi) (q^*(\pi))^\top \end{bmatrix}, \tag{126}$$

where $F_E(\cdot)$ is the matrix-valued function defined in (107), *i.e.*,

$$F_E(\chi_\pi) = \left(\frac{\tau}{1+\chi_\pi} E_{\text{tr}} + \pi B_{\text{test}} + \lambda\tau I_{d+1} \right)^{-1}. \tag{127}$$

Moreover,

$$q^*(\pi) = \frac{\tau}{(1+\chi_\pi)\sqrt{d}} \operatorname{vec}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} F_E(\chi_\pi) \right), \tag{128}$$

and

$$\frac{1}{c^*(\pi)} = \frac{\tau(1+\rho)}{1+\chi_\pi} + \lambda\tau - \frac{\tau^2}{(1+\chi_\pi)^2 d} \operatorname{tr}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} F_E(\chi_\pi) \begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix}^\top \right). \tag{129}$$

Observe that there is a one-to-one correspondence between the terms in (126) and those in (59).

To derive the asymptotic characterization given in (120), we note that

$$\begin{aligned}
\frac{1}{d} \operatorname{tr}(\Gamma^* A_{\text{test}}^\top) &\simeq \frac{-1}{c(0)\sqrt{d}} \begin{bmatrix} 0 & \operatorname{vec}(A_{\text{test}})^T \end{bmatrix} \mathcal{G}_e(0) e_1 && (130) \\
&= \frac{c^*(0)}{c(0)} \cdot \frac{1}{d} \operatorname{tr}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} (E_{\text{tr}} + \lambda(1+\chi_0) I)^{-1} A_{\text{test}}^\top \right) && (131) \\
&\simeq \frac{1}{d} \operatorname{tr}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho) b_{\text{tr}} \end{bmatrix} (E_{\text{tr}} + \lambda(1+\chi_0) I)^{-1} A_{\text{test}}^\top \right). && (132)
\end{aligned}$$

In the above display, (130) follows from (63) and the asymptotic equivalence between $G_e(0)$ and $\mathcal{G}_e(0)$. The equality in (131) is due to (126) and (128). To reach (132), we note that $c(0) = e_1^\top G_e(0) e_1$ and $c^*(0) = e_1^\top \mathcal{G}_e(0) e_1$. Thus, $c(0) \simeq c^*(0)$ due to the asymptotic equivalence between $G_e(0)$ and $\mathcal{G}_e(0)$. In Appendix B, we show that

$$\lambda(1+\chi_0) \simeq \xi, \tag{133}$$

where ξ is the scalar defined in (123). The asymptotic characterization given in (120) then follows from (132) and from the definition of Γ^*_{eq} given in (122).

Next, we use (65) to derive the asymptotic characterization of the quadratic term in (121). Taking the derivative of (129) gives us

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\pi} \left(\frac{1}{c^*(\pi)} \right) \bigg|_{\pi=0} &= \frac{1}{d} \operatorname{tr}(\Gamma^*_{\text{eq}} B_{\text{test}} (\Gamma^*_{\text{eq}})^\top) \\
&\quad - \frac{\tau\chi'_0}{(1+\chi_0)^2} \left(1 + \rho - \frac{2}{d} \operatorname{tr}(A_{\text{tr}} (E_{\text{tr}} + \xi I)^{-1} A_{\text{tr}}^T) + \frac{1}{d} \operatorname{tr}(A_{\text{tr}} (E_{\text{tr}} + \xi I)^{-1} E_{\text{tr}} (E_{\text{tr}} + \xi I)^{-1} A_{\text{tr}}^T) \right) && (134) \\
&= \frac{1}{d} \operatorname{tr}(\Gamma^*_{\text{eq}} B_{\text{test}} (\Gamma^*_{\text{eq}})^\top) - \frac{\tau\chi'_0}{(1+\chi_0)^2} \left(1 + \rho - \frac{1}{d} \operatorname{tr}(\Gamma^*_{\text{eq}} A_{\text{tr}}^T) - \frac{\xi}{d} \operatorname{tr}(\Gamma^*_{\text{eq}} (\Gamma^*_{\text{eq}})^\top) \right), && (135)
\end{aligned}$$

where A_{tr} is the matrix defined in (73). In reaching the above expression, we have also used the estimate in (133).

To further simplify our formula, we note that

$$A_{\text{tr}} = S\left(E_{\text{tr}} + \xi I_{d+1} - \left(\frac{1+\rho}{\alpha} + \xi\right) I_{d+1}\right), \tag{136}$$

where S is a $d \times (d+1)$ matrix obtained by removing the last row of I_{d+1}. Using this identity, we can rewrite the matrix Γ^*_{eq} in (122) as

$$\begin{aligned}\Gamma^*_{\text{eq}} &= S\left(I - \left(\frac{1+\rho}{\alpha} + \xi\right)(E_{\text{tr}} + \xi I)^{-1}\right) && (137)\\ &= \left[I - \nu F_R(\nu) - a^*(1+\rho)^2 \nu F_R(\nu) b_{\text{tr}} b_{\text{tr}}^\top F_R(\nu) \quad a^*(1+\rho)\nu F_R(\nu) b_{\text{tr}}\right], && (138)\end{aligned}$$

where $F_R(\cdot)$ is the function defined in (179), and ν is the parameter given in (125). The second equality (138) is obtained from the explicit formula for $(E_{\text{tr}} + \xi I)^{-1}$ in (185).

From (136) and (137), it is straightforward to check that

$$\frac{1}{d}\operatorname{tr}(\Gamma^*_{\text{eq}} A^T_{\text{tr}}) = 1 - \nu + \nu^2 \frac{1}{d}\operatorname{tr}(S(E_{\text{tr}} + \xi I)^{-1} S^\top), \tag{139}$$

and

$$\frac{\xi}{d}\operatorname{tr}(\Gamma^*_{\text{eq}}(\Gamma^*_{\text{eq}})^\top) = \xi\left[1 - 2\nu\frac{1}{d}\operatorname{tr}(S(E_{\text{tr}} + \xi I)^{-1} S^\top) + \nu^2 \frac{1}{d}\operatorname{tr}(S(E_{\text{tr}} + \xi I)^{-2} S^\top\right]. \tag{140}$$

By using the asymptotic characterizations given in (197) and (198), we then have

$$\frac{1}{d}\operatorname{tr}(\Gamma^*_{\text{eq}} A^T_{\text{tr}}) \simeq 1 - \nu + \nu^2 \mathcal{M}_\kappa(\nu), \tag{141}$$

and

$$\frac{\xi}{d}\operatorname{tr}(\Gamma^*_{\text{eq}}(\Gamma^*_{\text{eq}})^\top) \simeq \xi\left[1 - 2\nu \mathcal{M}_\kappa(\nu) - \nu^2 \mathcal{M}'_\kappa(\nu)\right]. \tag{142}$$

Substituting (141), (142), and (199) into (135) yields

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c^*(\pi)}\right)\bigg|_{\pi=0} \simeq \frac{1}{d}\operatorname{tr}(\Gamma^*_{\text{eq}} B_{\text{test}} (\Gamma^*_{\text{eq}})^T) - \frac{c_e}{d}\operatorname{tr}\left(B_{\text{test}}\left[(E_{\text{tr}} + \xi I)^{-1} - \xi(E_{\text{tr}} + \xi I)^{-2}\right]\right), \tag{143}$$

where c_e is the scalar defined in (124). The asymptotic characterization of the quadratic term in (121) then follows from (65) and the claim that

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c(\pi)}\right)\bigg|_{\pi=0} \simeq \frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c^*(\pi)}\right)\bigg|_{\pi=0}. \tag{144}$$

SI-13.2 The Generalization Error of In-Context Learning

Result 4. *Consider the test distribution $\mathcal{P}_{\text{test}}$ associated with the ICL task. We have*

$$e(\Gamma^*) \simeq e^{\text{ICL}}(\tau, \alpha, \kappa, \rho, \lambda), \tag{145}$$

where

$$\begin{aligned} e^{\text{ICL}}(\tau, \alpha, \kappa, \rho, \lambda) \coloneqq \left(\frac{1+\rho}{\alpha} + 1\right)\left(1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2\mathcal{M}'_\kappa(\nu) - c_e\left[\mathcal{M}_\kappa(\nu) + \xi\mathcal{M}'_\kappa(\nu)\right]\right)\\ - 2\left[1 - \nu\mathcal{M}_\kappa(\nu)\right] + 1 + \rho, \end{aligned} \tag{146}$$

and c_e is the constant given in (124).

Remark 6. *Recall the definition of the asymptotic equivalence notation "$\simeq$" introduced in Section SI-9. The characterization given in (145) implies that, as $d \to \infty$, the generalization error $e(\Gamma^*)$ converges almost surely to the deterministic quantity $e^{\text{ICL}}(\tau, \alpha, \kappa, \rho, \lambda)$.*

To derive (145), our starting point is the estimate

$$e(\Gamma^*) \simeq \frac{1}{d}\operatorname{tr}\left(\Gamma^* B_{\text{test}}(\Gamma^*)^\top\right) - \frac{2}{d}\operatorname{tr}\left(\Gamma^* A_{\text{test}}^\top\right) + 1 + \rho, \tag{147}$$

which follows from Corollary 1 and the discussions in Remark 2. We consider the ICL task here, and thus A_{test} and B_{test} are given in (119). The asymptotic limits of the first two terms on the right-hand side of the above equation can be obtained by the characterizations given in Result 3.

Using (120) and the expressions in (138) and (119), we have

$$\begin{aligned}\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\text{test}}^\top) &\simeq \frac{1}{d}\operatorname{tr}\left(\Gamma_{\text{eq}}^* A_{\text{test}}^\top\right) && (148)\\ &= 1 - \frac{\nu}{d}\operatorname{tr} F_R(\nu) - a^*(1+\rho)^2\nu\frac{\left\|F_R(\nu)b_{\text{tr}}\right\|^2}{d} && (149)\\ &\simeq 1 - \nu\mathcal{M}_\kappa(\nu), && (150)\end{aligned}$$

where ν is the constant defined in (125). To reach the last step, we have used the estimate given in (197).

Next, we use (121) to characterize the first term on the right-hand side of (147). From the formulas in (138) and (119), we can check that

$$\begin{aligned}\frac{1}{d}\operatorname{tr}\left(\Gamma_{\text{eq}}^* B_{\text{test}}(\Gamma_{\text{eq}}^*)^\top\right) &\simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr}\left(I - \nu F(\nu)\right)^2 && (151)\\ &\simeq \left(\frac{1+\rho}{\alpha}+1\right)\left(1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2\mathcal{M}_\kappa'(\nu)\right), && (152)\end{aligned}$$

where the second step follows from (197) and (198). From (185),

$$\frac{1}{d}\operatorname{tr}(B_{\text{test}}(E_{\text{tr}}+\xi I)^{-1}) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr} F_R(\nu) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\mathcal{M}_\kappa(\nu). \tag{153}$$

Similarly, we can check that

$$\frac{1}{d}\operatorname{tr}(B_{\text{test}}(E_{\text{tr}}+\xi I)^{-2}) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr} F_R^2(\nu) \simeq -\left(\frac{1+\rho}{\alpha}+1\right)\mathcal{M}_\kappa'(\nu). \tag{154}$$

Substituting (152), (153), and (154) into (121) gives us

$$\frac{1}{d}\operatorname{tr}(\Gamma^* B(\Gamma^*)^\top) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\left(1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2\mathcal{M}_\kappa'(\nu) - c_e\left[\mathcal{M}_\kappa(\nu) + \xi\mathcal{M}_\kappa'(\nu)\right]\right), \tag{155}$$

where c_e is the constant given in (124). Combining (150), (155), and (147), we are done.

In what follows, we further simplify the characterizations in Result 4 by considering the ridgeless limit, *i.e.*, when $\lambda \to 0^+$.

Result 5. *Let*

$$q^* := \frac{1+\rho}{\alpha}, \qquad m^* := \mathcal{M}_\kappa(q^*), \qquad \textit{and} \qquad \mu^* := q^*\mathcal{M}_{\kappa/\tau}(q^*), \tag{156}$$

where $\mathcal{M}_\kappa(x)$ is the function defined in (181). Then

$$\begin{aligned}e_{\text{ridgeless}}^{\text{ICL}} &:= \lim_{\lambda\to0^+} e^{\text{ICL}}(\tau,\alpha,\kappa,\rho,\lambda)\\ &= \begin{cases}\frac{\tau(1+q^*)}{1-\tau}\left[1-\tau(1-\mu^*)^2+\mu^*(\rho/q^*-1)\right] - 2\tau(1-\mu^*) + (1+\rho) & \tau<1\\ (q^*+1)\left(1-2q^*m^*-(q^*)^2\mathcal{M}_\kappa'(q^*)+\frac{(\rho+q^*-(q^*)^2m^*)m^*}{\tau-1}\right) - 2(1-q^*m^*) + (1+\rho) & \tau>1\end{cases},\end{aligned} \tag{157}$$

where $\mathcal{M}_\kappa'(\cdot)$ denotes the derivative of $\mathcal{M}_\kappa(x)$ with respect to x.

We start with the case of $\tau < 1$. Examining the self-consistent equation in (123), we can see that the parameter ξ tends to a nonzero constant, denoted by ξ^*, as $\lambda \to 0^+$. It follows that the original equation in (123) reduces to

$$\xi^*\mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi^*\right) = 1-\tau. \tag{158}$$

Introduce a change of variables

$$\mu^* := \frac{(1-\tau)(1+\rho)}{\alpha\tau\xi^*}. \tag{159}$$

By combining (158) and the characterization in (182), we can directly solve for μ and get $\mu^* = q^* \mathcal{M}_{\kappa/\tau}(q^*)$ as given in (156). The characterization in (157) (for the case of $\tau < 1$) then directly follows from (150), (155), and (3) after some lengthy calculations.

Next, we consider the case of $\tau > 1$. It is straightforward to verify from (123) that

$$\xi = \frac{\tau}{\tau - 1}\lambda + \mathcal{O}(\lambda^2). \tag{160}$$

Thus, when $\tau > 1$, $\xi \to 0$ as $\lambda \to 0^+$. It follows that

$$\lim_{\lambda \to 0^+} \nu = \lim_{\lambda \to 0^+} \left(\frac{1+\rho}{\alpha} + \xi \right) = q^* \quad \text{and} \quad \lim_{\lambda \to 0^+} \mathcal{M}_\kappa(\nu) = m^*. \tag{161}$$

Substituting these estimates into (150), (155), and (3), we then reach the characterizations in (157) for the case of $\tau > 1$.

A Equivalent Statistical Representations

In this appendix, we present an equivalent (but simplified) statistical model for the regression vector H_Z defined in (9). This statistically-equivalent model will simplify the moment calculations in Section SI-10 and the random matrix analysis in Section SI-12.

Lemma 3. *Let w be a given task vector with $\|w\| = \sqrt{d}$. Meanwhile, let $a \sim \mathcal{N}(0,1)$, $s \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,\rho)$ be three scalar normal random variables, and $q \sim \mathcal{N}(0, I_{\ell-1})$, $g \sim \mathcal{N}(0, I_{d-1})$, $u \sim \mathcal{N}(0, I_{d-1})$, and $v_\epsilon \sim \mathcal{N}(0, \rho I_\ell)$ be isotropic normal random vectors. Moreover, w and all of the above random variables are mutually independent. We have the following equivalent statistical representation of the pair $(H_Z, y_{\ell+1})$:*

$$H_Z \overset{(d)}{=} (d/\ell) M_w \begin{bmatrix} s \\ u \end{bmatrix} \begin{bmatrix} h^\top M_w, & (a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \end{bmatrix}, \tag{162}$$

and

$$y_{\ell+1} \overset{(d)}{=} s + \epsilon. \tag{163}$$

In the above displays, M_w denotes a symmetric *and* orthonormal *matrix such that*

$$(M_w) e_1 = \frac{w}{\|w\|}, \tag{164}$$

where e_1 denotes the first natural basis vector in $\mathbb{R}^d$; $h \in \mathbb{R}^d$ is a vector defined as

$$h := \begin{bmatrix} \frac{\theta_\epsilon a}{\sqrt{d}} + \frac{a^2}{d} + \theta_q^2 \\ \left[(\theta_\epsilon + a/\sqrt{d})^2 + \theta_q^2\right]^{1/2} g/\sqrt{d} \end{bmatrix}; \tag{165}$$

and θ_ϵ, θ_q are scalars such that

$$\theta_\epsilon = \|v_\epsilon\|/\sqrt{d} \qquad \text{and} \qquad \theta_q = \|q\|/\sqrt{d}. \tag{166}$$

Remark 7. *For two random variables A and B, the notation $A \overset{(d)}{=} B$ indicates that A and B have identical probability distributions. Note that A and B can be either scalars [as in the case of (163)], or matrices of matching dimensions [as in the case of (162)].*

Remark 8. *A concrete construction of the symmetric and orthonormal matrix M_w satisfying (164) can be based on the Householder transformation [42–44].*

Proof. Recall that the data vector $x_{\ell+1}$ is independent of the task vector w. Then, by the rotational symmetry of the isotropic normal distribution, we can rewrite

$$x_{\ell+1} \overset{(d)}{=} \frac{1}{\sqrt{d}} M_w \begin{bmatrix} s \\ u \end{bmatrix}, \tag{167}$$

where $s \sim \mathcal{N}(0,1)$ and $u \sim \mathcal{N}(0, I_{d-1})$ are two independent normal random variables (vectors), and M_w is the symmetric orthonormal matrix specified in (164). Note that $y_{\ell+1} = x_{\ell+1}^\top w + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \rho)$ denoting the noise. The representation in (163) then follows immediately from (167) and the identity in (164).

To show (162), we first reparameterize the $d \times \ell$ Gaussian data matrix X as

$$X = M_w \begin{bmatrix} a & q^\top \\ p & U \end{bmatrix} M_{v_\epsilon}/\sqrt{d}. \tag{168}$$

In the above display, $a \sim \mathcal{N}(0,1)$, $p \sim \mathcal{N}(0, I_{d-1})$, $q \sim \mathcal{N}(0, I_{\ell-1})$; $U \in \mathbb{R}^{(d-1)\times(\ell-1)}$ is a matrix with iid standard normal entries; and M_{v_ϵ} is a symmetric orthonormal matrix such that

$$M_{v_\epsilon} e_1 = \frac{v_\epsilon}{\|v_\epsilon\|}, \tag{169}$$

where e_1 denotes the first natural basis vector in $\mathbb{R}^\ell$. Since the data matrix X, the task vector w, and the noise vector v_ϵ are mutually independent, it is straightforward to verify via the rotational symmetry of the isotropic normal distribution that both sides of (168) have identical probability distributions. Using this new representation, we have

$$X v_\epsilon = \theta_\epsilon M_w \begin{bmatrix} a \\ p \end{bmatrix}. \tag{170}$$

Meanwhile,

$$X^\top w = M_{v_\epsilon} \begin{bmatrix} a \\ q \end{bmatrix}, \tag{171}$$

and thus

$$X X^\top w = \frac{1}{\sqrt{d}} M_w \begin{bmatrix} a^2 + \|q\|^2 \\ ap + Uq \end{bmatrix}. \tag{172}$$

Combining (171) and (172) yields

$$Xy = XX^\top w + X v_\epsilon \tag{173}$$

$$= M_w \begin{bmatrix} \theta_\epsilon a + a^2/\sqrt{d} + \theta_q^2 \sqrt{d} \\ (\theta_\epsilon + a/\sqrt{d})p + Uq/\sqrt{d} \end{bmatrix}. \tag{174}$$

Observe that $Uq/\sqrt{d} \overset{(d)}{=} \theta_q \, p'$, where $p' \sim \mathcal{N}(0, I_{d-1})$ is a normal random variable independent of everything else. Using this reparametrization for $Uq/\sqrt{d}$ and the fact that p, p' are two independent Gaussian vectors, we can conclude that

$$\frac{1}{\sqrt{d}} Xy \overset{(d)}{=} M_w h, \tag{175}$$

where h is the random vector defined in (165).

Lastly, we consider the term $y^\top y$ in (9). Since $y = X^\top w + v_\epsilon$,

$$y^\top y = \left\| X^\top w + v_\epsilon \right\|^2 \tag{176}$$

$$= \left\| X^\top w + \theta_\epsilon \sqrt{d} M_{v_\epsilon} e_1 \right\|^2 \tag{177}$$

$$= (a + \theta_\epsilon \sqrt{d})^2 + \theta_q^2 d, \tag{178}$$

where the second equality follows from (169) and to reach the last equality we have used the representation in (171). To show (162), we recall the definition of H_Z in (9). Substituting (167), (175) and (178) into (9), we are done. □

B The Stieltjes Transforms of Wishart Ensembles

In this appendix, we first recall several standard results related to the Stieltjes transforms of Wishart ensembles. In our problem, we assume that there are k unique task vectors $\{w_i\}_{i\in[k]}$ in the training set. Moreover, these task vectors $\{w_i\}_{i\in[k]}$ are independently sampled from the uniform distribution on the sphere $\mathcal{S}^{d-1}(\sqrt{d})$ with radius $\sqrt{d}$. Let

$$F_R(\nu) \coloneqq (R_{\mathrm{tr}} + \nu I_d)^{-1}, \tag{179}$$

where R_{tr} is the sample covariance matrix of the task vectors as defined in (72) and ν is a positive scalar.

Note that the distribution of R_{tr} is asymptotically equivalent to that of a Wishart ensemble. By standard random matrix results on the Stieltjes transforms of Wishart ensembles (see, *e.g.*, [24]), we have

$$\frac{1}{d}\operatorname{tr} F_R(\nu) \simeq \mathcal{M}_\kappa(\nu) \tag{180}$$

as $d, k \to \infty$ with $k/d = \kappa$. Here,

$$\mathcal{M}_\kappa(\nu) \coloneqq \frac{2}{\nu + 1 - 1/\kappa + \left[(\nu + 1 - 1/\kappa)^2 + 4\nu/\kappa\right]^{1/2}}. \tag{181}$$

is the solution to the self-consistent equation

$$\frac{1}{\mathcal{M}_\kappa(\nu)} = \frac{1}{1 + \mathcal{M}_\kappa(\nu)/\kappa} + \nu. \tag{182}$$

Moreover,

$$\frac{1}{d}\operatorname{tr} F^2(\nu) \simeq -\mathcal{M}'_\kappa(\nu) = \frac{\mathcal{M}^2_\kappa(\nu)}{1 - \frac{\kappa \mathcal{M}^2_\kappa(\nu)}{[\kappa + \mathcal{M}_\kappa(\nu)]^2}}. \tag{183}$$

For the remainder of this appendix, we will further explore the self-consistent equation given by (78). We will show that the solution χ_π and its derivative $\frac{d}{d\pi}\chi_\pi$, at $\pi = 0$, can be characterized by the function $\mathcal{M}_\kappa(\nu)$ in (181). To start, note that at $\pi = 0$, the equation in (78) can be written as

$$\frac{\tau\chi_0}{1+\chi_0} = (1 + 1/d) - \frac{\lambda(1+\chi_0)}{d}\operatorname{tr}(E_{\mathrm{tr}} + \lambda(1+\chi_0)I)^{-1}. \tag{184}$$

Recall the definition of E_{tr} given in (74). It is straightforward to verify that

$$(E_{\mathrm{tr}} + \lambda(1+\chi_0)I_{d+1})^{-1} = \begin{bmatrix} F_R(\nu_0) + a^*(1+\rho)^2 F_R(\nu_0) b_{\mathrm{tr}} b_{\mathrm{tr}}^\top F_R(\nu_0) & -a^*(1+\rho)F_R(\nu_0)b_{\mathrm{tr}} \\ -a^*(1+\rho) b_{\mathrm{tr}}^\top F_R(\nu_0) & a^* \end{bmatrix}, \tag{185}$$

where $F_R(\cdot)$ is the function defined in (179),

$$\nu_0 = \frac{1+\rho}{\alpha} + \lambda(1+\chi_0) \tag{186}$$

and

$$\frac{1}{a^*} = (1+\rho)^2 + \lambda(1+\chi_0) - (1+\rho)^2 b_{\mathrm{tr}}^\top F_R(\nu_0) b_{\mathrm{tr}}. \tag{187}$$

From (185), the equation (184) becomes

$$\frac{\tau\chi_0}{1+\chi_0} = (1 + 1/d) - \frac{\lambda(1+\chi_0)}{d}\operatorname{tr} F_R(\nu_0) - (1+\rho)^2 \frac{a^*\lambda(1+\chi_0)}{d}\left\|F_R(\nu_0) b_{\mathrm{tr}}\right\|^2. \tag{188}$$

By the construction of $F_R(\nu_0)$ and b_{tr}, we can verify that

$$b_{\mathrm{tr}}^\top F_R(\nu_0) b_{\mathrm{tr}} \le 1 \qquad \text{and} \qquad \left\|F_R(\nu_0) b_{\mathrm{tr}}\right\|^2 \le \frac{1}{\nu_0} \le \frac{\alpha}{1+\rho}. \tag{189}$$

Substituting the first inequality above into (187) gives us

$$a^*\lambda(1+\chi_0) \le 1. \tag{190}$$

Combining this estimate with the second inequality in (189), we can conclude that the last term on the right-hand side of (188) is negligible as $d \to \infty$. Moreover, using the asymptotic characterization given in (180), the equation (188) leads to

$$\frac{\tau \chi_0}{1+\chi_0} \simeq 1 - \lambda(1+\chi_0)\mathcal{M}_\kappa(\nu_0). \tag{191}$$

Introducing a change of variables

$$\xi_0 = \lambda(1+\chi_0), \tag{192}$$

and also recalling the definition of ν_0 in (186), we can further transform (191) to

$$\xi_0 \mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha} + \xi_0\right) - \frac{\tau\lambda}{\xi_0} \simeq 1 - \tau. \tag{193}$$

Observe that the above is identical to the equation in (123), except for a small error term captured by $\simeq$. By the stability of (123), we can then conclude that

$$\xi_0 \simeq \xi, \tag{194}$$

thus verifying (133).

Next, we compute χ_0', the derivative of χ_π (with respect to π) evaluated at $\pi = 0$. Differentiating (78) give us

$$\tau\chi_0' = \frac{1}{d}\operatorname{tr}\left[(E_{\text{tr}} + \xi_0 I)^{-1}\Big(\chi_0' E_{\text{tr}} - \frac{(1+\chi_0)^2}{\tau} B_{\text{test}}\Big)(E_{\text{tr}} + \xi_0 I)^{-1} E_{\text{tr}}\right]. \tag{195}$$

Thus,

$$\frac{\tau\chi_0'}{(1+\chi_0)^2} \simeq \frac{\frac{1}{d}\operatorname{tr}\left(B_{\text{test}}[(E_{\text{tr}}+\xi I)^{-1} - \xi(E_{\text{tr}}+\xi I)^{-2}]\right)}{1 - 2\xi\operatorname{tr}(E_{\text{tr}}+\xi I)^{-1}/d + \xi^2\operatorname{tr}(E_{\text{tr}}+\xi I)^{-2}/d - \tau}, \tag{196}$$

where we have used (194) to replace ξ_0 in (195) by ξ, with the latter being the solution to the self-consistent equation in (123). Using the decomposition in (185) and following similar arguments that allowed us to simplify (188) to (191), we can check that

$$\frac{1}{d}\operatorname{tr}(E_{\text{tr}}+\xi I)^{-1} \simeq \frac{1}{d}\operatorname{tr} S(E_{\text{tr}}+\xi I)^{-1}S^\top \simeq \frac{1}{d}\operatorname{tr} F\left(\frac{1+\rho}{\alpha}+\xi\right) \simeq \mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right), \tag{197}$$

and

$$\frac{1}{d}\operatorname{tr}(E_{\text{tr}}+\xi I)^{-2} \simeq \frac{1}{d}\operatorname{tr} S(E_{\text{tr}}+\xi I)^{-2}S^\top \simeq \frac{1}{d}\operatorname{tr} F^2\left(\frac{1+\rho}{\alpha}+\xi\right) \simeq -\mathcal{M}_\kappa'\left(\frac{1+\rho}{\alpha}+\xi\right), \tag{198}$$

where S is a $d \times (d+1)$ matrix obtained by removing the last row of I_{d+1}, and $\mathcal{M}_\kappa(\cdot)$ is the function defined in (181). Substituting (197) and (198) into (196) yields

$$\frac{\tau\chi_0'}{(1+\chi_0)^2} \simeq \frac{\frac{1}{d}\operatorname{tr}\left(B_{\text{test}}[(E_{\text{tr}}+\xi I)^{-1} - \xi(E_{\text{tr}}+\xi I)^{-2}]\right)}{1 - 2\xi\mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right) - \xi^2\mathcal{M}_\kappa'\left(\frac{1+\rho}{\alpha}+\xi\right) - \tau}. \tag{199}$$